The Dissertation Committee for Jing Zan
certifies that this is the approved version of the following dissertation:

Staffing Service Centers Under Arrival-rate Uncertainty

Committee:

John J. Hasenbein, Supervisor

David P. Morton, Supervisor

J. Eric Bickel

Qi Feng

Elmira Popova

Staffing Service Centers Under Arrival-rate Uncertainty

by

Jing Zan, B.S., M.S.E.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2012

To everyone who has helped and supported me over the years.

Acknowledgments

First and foremost, I would like to thank my advisors, Dr. John Hasenbein and Dr. David Morton for their guidance in my research work. I greatly appreciate their stimulation, optimism, encouragement, and patience in my work, without which I would not have been able to make as much progress as I have achieved. Their enthusiasm for research always brings brilliant new ideas to move our projects forward, and at the same time, they give me a lot of freedom to explore the topics that I am most interested in. I am also very grateful to them for helping me improve my English speaking and writing skills. Working with John and David has been fun and I have learned a lot from them.

I would also like to express my gratitude to my other committee members Dr. Popova, Dr. Bickel and Dr. Feng for agreeing to be my committee members, providing valuable comments for my research work, teaching me optimization, statistics, simulation and modeling concepts and techniques, and furthermore sharing their study and research experience with me, which are especially important to me during the tough time of my doctoral studies.

I owe many thanks to Dr. Vijay Mehrotra from University of San Francisco for sharing his industrial experiences in call center industry with me. After discussing with him, I form a clearer idea on the practical needs of the

call center industry. The research topic in Chapter 5 is also chosen with Dr. Mehrotra's help.

Finally, thanks are due to my family for their understanding, support, and patience throughout these years, which has been a constant source of courage to keep me going.

Staffing Service Centers Under Arrival-rate Uncertainty

Publication No. _____

Jing Zan, Ph.D.
The University of Texas at Austin, 2012

Supervisors: John J. Hasenbein
David P. Morton

We consider the problem of staffing large-scale service centers with multiple customer classes and agent types operating under quality-of-service (QoS) constraints. We introduce formulations for a class of staffing problems, minimizing the cost of staffing while requiring that the long-run average QoS achieves a certain pre-specified level. The queueing models we use to define such service center staffing problems have random inter-arrival times and random service times. The models we study differ with respect to whether the *arrival rates* are deterministic or stochastic. In the deterministic version of the service center staffing problem, we assume that the customer arrival rates are known deterministically.

It is computationally challenging to solve our service center staffing problem with deterministic arrival rates. Thus, we provide an approximation and prove that the solution of our approximation is asymptotically optimal in the sense that the gap between the optimal value of the exact model and the

objective function value of the approximate solution shrinks to zero as the size of the system grows large.

In our work, we also focus on doubly stochastic service center systems; that is, we focus on solving large-scale service center staffing problems when the arrival rates are uncertain in addition to the inherent randomness of the system's inter-arrival times and service times. This brings the modeling closer to reality. In solving the service center staffing problems with deterministic arrival rates, we provide a solution procedure for solving staffing problems for doubly stochastic service center systems. We consider a decision making scheme in which we must select staffing levels before observing the arrival rates. We assume that the decision maker has distributional information about the arrival rates at the time of decision making. In the presence of arrival-rate uncertainty, the decision maker's goal is to minimize the staffing cost, while ensuring the QoS achieves a given level. We show that as the system scales large in size, there is at most one *key* scenario under which the probability of waiting converges to a non-trivial value, i.e., a value strictly between 0 and 1. That is, the system is either over- or under-loaded in any other scenario as the size of the system grows to infinity. Exploiting this result, we propose a two-step solution procedure for the staffing problem with random arrival rates. In the first step, we use the desired QoS level to identify the key scenario corresponding to the optimal staffing level. After finding the key scenario, the random arrival-rate model reduces to a deterministic arrival-rate model. In the second step, we solve the resulting model, with deterministic arrival

rate, by using the approximation model we point to above. The approximate optimal staffing level obtained in this procedure asymptotically converges to the true optimal staffing level for the random arrival-rate problem.

The decision making scheme we sketch above, assumes that the distribution of the random arrival rates is known at the time of decision making. In reality this distribution must be estimated based on historical data and experience, and needs to be updated as new observations arrive. Another important issue that arises in service center management is that in the daily operation in service centers, the daily operational period is split into small decision time periods, for example, hourly periods, and then the staffing decisions need to be made for all such time periods. Thus, to achieve an overall optimal daily staffing policy, one must deal with the interaction among staffing decisions over adjacent time periods. In our work, we also build a model that handles the above two issues. We build a two-stage stochastic model with recourse that provides the staffing decisions over two adjacent decision time periods, i.e., two adjacent decision stages. The model minimizes the first stage staffing cost and the expected second stage staffing cost while satisfying a service quality constraint on the second stage operation. A Bayesian update is used to obtain the second-stage arrival-rate distribution based on the first-stage arrival-rate distribution and the arrival observations in the first stage. The second-stage distribution is used in the constraint on the second stage service quality. After reformulation, we show that our two-stage model can be expressed as a newsvendor model, albeit with a demand that is derived from the first stage

decision. We provide an algorithm that can solve the two-stage staffing problem under the most commonly used QoS constraints.

This work uses stochastic programming methods to solve problems arising in queueing networks. We hope that the ideas that we put forward in this dissertation lead to other attempts to deal with decision making under uncertainty for queueing systems that combine techniques from stochastic programming and analysis tools from queueing theory.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation: Service Center Systems

Service centers, which handle more than 70% (Borst et al. [9]) of the customer-business interactions in the US, have been viewed as the modern business frontier. With estimated annual expenditures exceeding $300 billion (Gilson and Khandelwal [19]), the service center industry has received increased attention from both business and from the operations research community. Although, with the development of new technologies, the service center business has become more technology-intensive, most of its operating costs are still devoted to human resources. Recent statistics show that 3% (Borst et al. [9]) of the US workforce is devoted to the service center industry, with an annual rate of growth of more than 8% (Gans et al. [17]). With 60-80% (Aksin et al. [1]) of the service center operating costs coming from labor costs, service center managers are tempted to reduce the number of servers so as to cut labor costs. However, doing so may risk quality of service, such as making customers wait too long, causing them to either abandon the system or fume over poor service. Such outcomes may incur penalty costs (for third-party providers) or damage the corporate image. This naturally gives rise to an interest in finding an optimal staffing policy to attain the desired trade-off

between service quality and operational efficiency.

A service center is a place used to respond to requests from customers. It usually consists of a set of resources such as staff, computers, and telecommunication equipment, which enable the delivery of services via telephone or other customer interface. Most organizations today have service centers, either internally managed or outsourced, as their primary customer-facing channel. According to [14], there are over 80,000 call centers in the United States.

Service centers can differ in the types of services they provide. Some service centers may only offer one type of service, but mostly, a single service center provides highly varied services, such as technical support, new product introduction, customer sales, etc. A service center that offers only one type of service is a single-class, single-station system, in the sense that there is only one customer class and one type of service station. The staffing decision for a single-class, single-station system is relatively easier to make than that for more complicated service center systems. A service center that offers more than one type of service is a multi-class, multi-station system. For such a system, customers with different kinds of requests arrive to the system, and for each type of request, there may be a particular service station that has been designed to respond to that class of request. The operations on multi-class, multi-station systems can be very complicated in the sense that some of the service agents may be cross-trained to handle multiple types of requests, and the arrival of different types of customers may be correlated. In such a situation, in making the decision on the staffing levels of all service

stations, the service center manager may need to take the interaction among arrivals into consideration, and after making the staffing decision, the manager is confronted with a scheduling problem involving that staff. We consider both single-class, single-station systems and multi-class, multi-station systems. And, we focus on staffing issues as opposed to scheduling.

Service centers can vary considerably in size, from small size with a few service agents, like a hotel front desk, to a huge size service center, like a huge call center embodying several smaller call centers across different geographical locations. In our work, we focus on large service centers. The service quality and operational efficiency can be relatively high in well-managed large service centers. Many hundreds of service agents can handle many thousands of service requests per hour.

1.2 Relevant Queueing Literature

Service center systems are stochastic systems because they contain random elements like: the arrival of customers; the time it takes agents to serve customers; and, the time before a customer abandons the system. Queueing models are usually used to represent such stochastic systems. The queueing models describing call center operations can be complicated, especially when taking into account factors such as the behavior of customers in abandoning the system, the effect of different levels of training of servers, etc. The most basic and widely-used queueing model for call centers is the $M/M/n$ system, also known as the Erlang-C system (Erlang [15, 16]). The $M/M/n$ system

describes a single-class, single-station service center. It assumes a steady-state environment in which arrivals conform to a Poisson process, service durations are exponentially distributed, and servers are statistically identical and work independently of each other. The $M/M/n$ system neglects the issue of customer abandonment. However, even the most basic and analytically solvable model, the Erlang-C model, is not completely useful in practical application, because of the computational intractability of the expression for its steady-state distribution. For this reason, there has been significant research interest in developing approximations of the queueing model, in large part by applying asymptotic analysis. Another reason for the research interest in developing approximations of a queueing model is that, instead of obtaining a staffing level decision for the service center at a specific arrival rate, it can be valuable to gain insight regarding structural properties of the solutions to service center problems. Proper approximation may help to understand such structures, rather than just providing an approximate solution.

The bulk of the approximations of the $M/M/n$ queue in the literature become more accurate as the arrival rate grows large. As a result, under appropriate conditions, the associated asymptotic analysis provides near optimal policies when the arrival rates are sufficiently large. Halfin and Whitt [22] propose an asymptotic regime for analyzing a heavy-traffic Erlang-C system. They consider a sequence of systems in which service rates are fixed, arrival rates and the number of servers go to infinity, and the utilization goes to 1 from below. This scaling is designed to keep the probability that a cus-

tomer waits for service close to a predetermined value strictly between 0 and 1. Using this "Halfin-Whitt approximation" the probability that a customer waits is much simpler to compute, as opposed to using the exact Erlang-C formula. The so-called square-root staffing policy is derived from the Halfin and Whitt result, and it is shown that when the arrival rate is large, we can achieve high utilization and provide a good level of service by applying the square-root staffing policy. Halfin and Whitt also show that as the arrival rate grows toward infinity, the probability of waiting converges to some value strictly between 0 and 1 if and only if the square-root staffing policy is applied.

Numerous recent papers examine systems in the so-called Halfin-Whitt regime. Armony [3] and Gurvich et al. [20], for example, examine staffing and routing problems in this regime. Based on Halfin and Whitt's work, efforts have been devoted to obtain better approximations. Janssen et al. [25] derive a refinement of the Halfin and Whitt approximation, which is more accurate while still simple to compute. They also provide theoretical support for the idea that a square-root safety staffing policy works well for moderate-size systems. Borst et al. [9] also revisit the square-root safety staffing principle and develop a new framework for asymptotic optimization of an $M/M/n$ system. In [25] and [9], similar to our work, the approximations for steady-state queueing systems are applied, but the model can only handle a single-class and single-station service system. Atar [4] considers the staffing problem for multi-class service system with heterogeneous service stations. Atar builds a diffusion model and constructs solutions based on solutions to a system of

partial differential equations. Because of the complexity in solving the partial differential equations, the method is computationally intractable for large-size problems.

In our work, we propose a computationally tractable model using an approximation for steady-state queueing measures, which can be applied to multi-class and multi-station service systems. In such a queueing system, while the inter-arrivals are described as exponentially distributed random variables, the overall arrival rate is assumed to be known and certain. However, this is arguably unrealistic in practice, where the arrival rate must be estimated based on experience and historical data. It is risky to ignore the randomness in the arrival rate. For example, suppose the staffing decision is made under a fixed chosen arrival rate, but the fact is that half of the time, the arrival rate will take the fixed chosen value, and the other half of the time, the actual arrival rate will take a value twice as large. Then, half of the time the queueing system may not have a steady state and the waiting times will drift off to infinity.

In recent work that we review in detail below, researchers have begun to realize the importance of incorporating arrival-rate randomness into the model formulations. Several papers tackle the problem of parameter randomness by formulating specific forecasting models (e.g., Brown and Shen [11], Brown et al. [10], Shen and Huang [34]). They make use of statistical analysis to form mathematical models that depict the characteristics for call center parameters.

Heavy-traffic approximations are used in some studies on call centers with random parameters, especially for single-class, single-type call centers.

Maman [27] achieves an asymptotically optimal staffing policy, that minimizes the staffing level while satisfying certain cost constraints for an $M/M/n+M$ queue with uncertain arrival rates, by using many-server, heavy-traffic approximations based on expected arrival rates.

To reduce the difficulty in handling doubly stochastic systems, fluid models are useful tools. We can view a fluid model as a deterministic counterpart to the original queueing network. Using a fluid model to approximate an original queueing system, can help to eliminate short-time-scale randomness in these systems, so as to reduce the difficulty in analyzing or solving the associated problem.

Harrison and Zeevi [23] solve the staffing problem for a multi-class, multi-server call center with random arrival rates by using a fluid model to approximate the system. The fluid approximation is a continuous state system which involves only the mean values of the original stochastic system. The authors build a model with the objective of minimizing the sum of the staffing cost and the penalty cost associated with the abandonment, and seek to deal with the trade-off between hiring too many servers and inducing a large abandonment penalty. Their proposed two-stage stochastic program with recourse incorporates arrival rate uncertainty into the staff scheduling step, by combining the first stage staff scheduling problem and the second stage dynamic routing problem. Robbins and Harrison [33] also build a stochastic programming model for determining staffing levels in call centers that are under a service level constraint with uncertain arrival rates. They conduct experi-

ments to show that the result achieved from solving their model is better than the result achieved from solving a deterministic model with expected arrival rates.

Bassamboo et al. [6] consider a fluid approximation for multi-class, multi-type call centers. They use a linear-programming based method to solve for an asymptotically optimal staffing and routing policy that minimizes the staffing cost and abandonment penalty. Bassamboo and Zeevi [7] extend the work in [6], using a data-driven method that provides the optimal staffing level without knowing the probabilistic structure of the arrival rates.

Also using a fluid approximation, Gurvich et al. [21] build a chance-constrained formulation, which yields the staffing and routing policy for multi-class, multi-type call centers with arrival rate uncertainty. In their work, they first find a small set of arrival-rate vectors and then perform a simulation-based search on the set of arrival-rate vectors found in the first step for the optimal staffing solution. Their approach deals with the uncertainty in arrival rates by translating the problem with uncertain arrival rates to a set of problems with perfectly known arrival rates. Their procedure provides a feasible solution that is nearly optimal. Fluid models may only work well for very large service centers in which the arrival-rate uncertainty is significant.

In our work, we do not appeal to fluid limits and instead employ queueing models with full short-time-scale stochastic dynamics to formulate service center staffing problems. Furthermore, unlike previous related work, our models can also handle multi-class, multi-station service center problems.

In the models we sketch above, a single staffing decision is made in a model that, once the arrival rate uncertainty is revealed, operates in steady state. Such a model does not account for the level of adaptivity that exists in some systems. For this reason, we also introduce a model that allows us to adjust staffing levels in, say, two adjacent four-hour time stages. In doing so, it is important to capture costs incurred for increasing or decreasing the level of the workforce over these time scales. And, it is important to use a probability model that uses arrival observations in the first stage in order to update our estimate of the second stage arrival distribution.

In our work, we apply stochastic programming with recourse to model this problem. Robbins and Harrison [33] formulate a service center problem as a two-stage mixed integer stochastic program to combine the staffing policy and staff scheduling decision, the decision made after the staffing policy is made, into a single optimization program. They also show that the results from the stochastic formulation significantly reduce the expected cost of operation, compared to the results from a deterministic program based on mean valued arrivals. Gans et al. [18] propose an approach to include arrival-rate updates, again using a two-stage stochastic program with recourse. In both [18] and [33], the authors focus on the service center scheduling problem or the scheduling problem nested within the staffing problem. In our work, we focus on staffing decisions over two adjacent time periods using a two-stage stochastic program with recourse. Stochastic programming with recourse has been applied to other problems of a similar nature as service center problems. Büke et al.

[13] develop a multi-stage stochastic fluid model in analyzing the makespan problem in semiconductor wafer fabrication scheduling.

1.3 Relevant Forecasting Literature

To model service center problems, there are certain parameters that need to be known. For example, the Erlang-C model requires arrival rates and service rates for the associated inter-arrival and service-time exponential distributions. Usually, these parameters are estimated by combining statistical analysis and empirical research. In our work, we largely focus on arrival rates. Bianchi et al. [8] and Andrews and Cunningham [2] apply autoregressive moving-average (ARMA) models to describe the arrival process. Shen and Huang [34] use singular value decomposition to analyze arrival data. A dynamic harmonic regression model for service center hourly arrivals is proposed in Tych et al. [41] that behaves better than seasonal ARMA models. Other work uses a non-homogeneous Poisson process to capture the time-varying property of the arrival rate. Weinberg et al. [42] provide a multiplicative Gaussian model for predicting the arrival rates of a non-homogeneous Poisson process. They use Bayesian procedures to fit their model and provide not only point estimates on the arrival rates, but also the probability distribution of the arrival rates. Jongbloed and Koole [26] handle the overdispersion in the arrivals relative to the Poisson distribution by proposing a doubly stochastic model. They assume the arrivals follow a Poisson process with gamma distributed arrival rate. In modeling the time-varying arrival rate in their ap-

plications, they assume the arrivals in different time periods within the same day are independent. That is they assume zero correlations between the arrival counts in different time periods. Brown et al. [10] and Avramidis et al. [5] incorporate inter- and intra-day correlation and dependence structures into their forecasting models for a non-homogenous Poisson arrival process.

As researchers began to realize the effect of the randomness in the arrival-rate, more work has focused on building models to forecast random arrival rates. Shen and Huang [35] introduce a dynamic factor model for a Poisson arrival process with random rate. Soyer and Tarimcilar [38] provide a modulated Poisson process model with stochastic arrival rates for describing the arrival process and again employ Bayesian analysis, using Gibbs sampling, to estimate associated parameters. Shen et al. [37] develop a Poisson factor model that combines the data-driven approach in Shen and Huang [36] and the model-driven approach in Weinberg et al. [42] to provide inter-day forecasting and intra-day updating. Taylor [40] adapts Holt-Winters exponential smoothing for modeling both the intra-day and intra-week cycles in intra-day data, and his method performs well compared to an ARMA model for lead times up to about four days ahead. Taylor [39] further develops the method in Taylor [40] to enable probability density function forecasting of both the number of arrivals and the inter-arrival rate for intra-day data. He develops a Poisson count data model, which captures the important features of the method developed in Taylor [40], with a gamma distributed stochastic arrival rate. In Taylor [39], a call center simulation model is used to demonstrate the

relationship between density forecasting and operational decision-making.

1.4 Dissertation Organization

The remaining of this dissertation is organized as follows. Chapter 2 introduces the mathematical background for our work. We review two widely applied queueing models for service center problems, the Erlang-C model and the Erlang-A model. We also detail our mathematical results on the approximations of the Erang-C model, which we use in later chapters. In Chapter 3, we present an approximation model for the situation in which the arrival rate is know. We prove asymptotic optimality of solutions for our approximation model for both a single-class single-station system and a multi-class multi-station system. In Chapter 4, we extend our approximation model in Chapter 3 for the setting in which we have stochastic arrival rates. We propose a two-step solution procedure for solving the staffing problem with random arrival rates and prove that our solution is asymptotically optimal. We formulate a two-stage stochastic program with recourse in Chapter 5 to analyze the relationship between the staffing decisions over two adjacent time periods. This chapter focuses on problems with a single class and a single station, under a random arrival rate. A Bayesian update is applied to the arrival rate in the second time period, after we observe arrivals in the first period. The model integrates arrival-rate updates and dependence in staffing decisions over two contiguous time periods. Chapter 6 summarizes the contributions of the dissertation and discusses future work.

Chapter 2

Mathematical Background

In this chapter, we introduce the mathematical background for our work. We review two widely applied queueing models for service center problems, the Erlang-C model and the Erlang-A model, which are the basic models we use in our work. We also introduce the approximations for the Erlang-C formula applied to modeling the staffing problems in our work. The mathematical proofs for key properties, such as uniform convergence of the approximations to the actual formula, are given in this chapter. Such properties of the approximations play a key role in proving the optimality of our staffing policies in the later chapters.

We first introduce some common mathematical notation used in our work. In our work, we use $\mathbb{N}$ stand for the natural numbers $\{1, 2, \ldots, \}$, $\mathbb{Z}_+$ stands for the non-negative integer numbers $\{0, 1, 2, \ldots, \}$, and $\mathbb{R}_+$ stands for the non-negative real numbers. We use $\phi(\cdot)$ to represent the probability density function (PDF) of a standard normal distribution and $\Phi(\cdot)$ to represent the cumulative distribution function (CDF) of a standard normal distribution. In our work $\Gamma(\cdot)$ stands for the gamma function, $\Gamma(\cdot, \cdot)$ stands for the upper incomplete gamma function and $\gamma(\cdot, \cdot)$ stands for the lower incomplete gamma

function.

2.1 $M/M/n$ Queue

2.1.1 Erlang-C Formula

The $M/M/n$ system is an easy system to analyze. Although it is a relatively simple model, it is a good starting point for analyzing more complicated systems. The $M/M/n$ system, or the so-called Erlang-C system, models a call center with n servers. The arrival of calls follows a Poisson process with rate λ customers per unit of time, and service times are exponentially distributed with mean $1/\mu$ units of time. Without loss of generality, in our work, we use minute as the unit of time. The service times are independent of each other and of the arrival process. The Erlang-C queueing model is shown in Figure 2.1. For simplicity, without loss of generality, we assume $\mu = 1$. An arriving customer entering the system receives service immediately if there is an idle server, or, if all servers are busy, the customer waits in the queue until a server is free. The service discipline is assumed to be first-come-first-served (FCFS).

The number of customers in the system forms a birth-death process, and its steady-state distribution can be obtained using standard theory. The steady-state number of customers in the system, denoted Q, has the following distribution:

$$\mathbb{P}\{Q = k\} = \begin{cases} \eta \frac{\lambda^k}{k!} & \text{if } k < n, \\ \eta \frac{n^n (\lambda/n)^k}{n!} & \text{otherwise,} \end{cases}$$

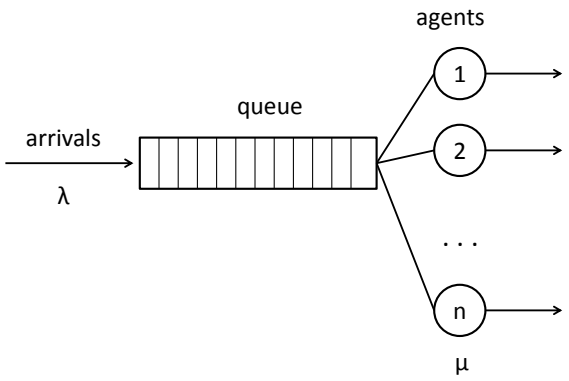

Figure 2.1: Erlang-C Model

where η is a normalizing constant,

$$\eta = \left[\sum_{k=0}^{n-1} \frac{(\lambda)^k}{k!} + \frac{(\lambda)^n}{n!(1-\lambda/n)} \right]^{-1}.$$

The above result allows us to compute $\mathbb{P}\{Q \geq n\}$, which, according to the property that Poisson arrivals see time averages (PASTA), is equal to the stationary probability that a customer waits to receive service. The so-called Erlang-C formula is used to calculate this probability of waiting, denoted as α:

$$\alpha(n, \lambda) = \eta \frac{(\lambda)^n}{n!(1-\lambda/n)}.$$

We now introduce a continuous extension of the Erlang-C formula (Jagers and Van Doorn [24]). This continuous Erlang-C formula, which al-

lows the number of servers n to take a non-integer value, gives the same value as the Erlang-C formula for the integer points of n, and is defined as follows:

$$\tilde{\alpha}(n,\lambda) = \left[\lambda \int_0^\infty te^{-\lambda t}(1+t)^{n-1}dt\right]^{-1}. \tag{2.1}$$

The continuous version of the Erlang-C formula allows us to consider the staffing problem with continuous decision variables, which facilitates some of our analysis. In our work, we build models for service center staffing problems with continuous decision variables.

2.1.2 Halfin-Whitt Approximation

For the $M/M/n$ system, we have the stationary distribution for the number of customers in the system and can apply the Erlang-C formula to obtain the probability that a customer waits for service, which is an indicator of service quality. In practice, however, many call centers are large, and the Erlang-C formula can become burdensome, and even intractable, for computations with large systems. Furthermore, we have interest in obtaining structural insights into the system. For example, it is obvious that for an $M/M/n$ system, as the arrival rate goes to infinity, the number of servers must also go to infinity, at least if we wish to maintain a stable system, or maintain a certain QoS level. More importantly, we need to know specifically how the number of servers and the arrival rate should go to infinity together. To achieve our goal, we carry out an asymptotic analysis and obtain an asymptotic solution which is increasingly accurate as the arrival rate grows large. Through this

asymptotic analysis, we can build a more tractable approximation of the original complicated queueing system, whose accuracy improves as the size of the system, measured by the number of servers, increases.

To carry out an asymptotic analysis, Halfin and Whitt [22] consider a sequence of $M/M/n$ queues, indexed by n. As n increases the scaling of λ is such that the server utilization $\rho_n := \frac{\lambda_n}{n\mu}$ approaches 1 while $\rho_n < 1$ for each n. As indicated above, to simplify notation, we assume the service rate, μ, to be 1. Such sequences lie in the so-called heavy-traffic regime. For our problem, we want to find a heavy-traffic regime where α_n, the probability of waiting, converges to an α that is strictly between 0 and 1. Halfin and Whitt prove that only sequences for which $\sqrt{n}(1-\rho_n)$ converges to a finite positive constant lead to an α between 0 and 1. For completeness we now restate their classic result.

Theorem 1. *(Halfin and Whitt [22])* As $n \to \infty$, the probability of waiting α_n converges to α with $0 < \alpha < 1$ if and only if

$$\sqrt{n}(1-\rho_n) \to \beta \tag{2.2}$$

for some $\beta > 0$. If (2.2) holds, then

$$\alpha = \frac{1}{1 + \sqrt{2\pi}\beta\Phi(\beta)e^{\beta^2/2}}. \tag{2.3}$$

Theorem 1 implies that when the system is large enough, (2.3), which is called the Halfin-Whitt approximation, approximates the Erlang-C formula well.

2.1.3 Bounds for Erlang-C Formula

Besides the well-known Halfin and Whitt approximation, there are other good approximations of the Erlang-C formula. Janssen et al. [25] provide useful bounds for the Erlang-C formula which have a simpler structure than the Erlang-C formula itself and hold for the continuous version of the Erlang-C formula. In our work, we use the bounds provides by Janssen et al. to build our approximate model for the service center staffing problem and achieve our approximate optimal solutions.

Theorem 2. *(Janssen et al. [25])* Let $\rho = \lambda/n$ and

$$\alpha = \sqrt{-2n(1-\rho+\ln\rho)}, \tag{2.4}$$

$$\beta = (n-\lambda)/\sqrt{\lambda}, \tag{2.5}$$

$$\gamma = (n-\lambda)/\sqrt{n} = \beta\sqrt{\rho}. \tag{2.6}$$

For $n > \lambda$,

$$\tilde{\alpha}(n,\lambda) \leq \left[\rho + \gamma\left(\frac{\Phi(\alpha)}{\phi(\alpha)} + \frac{2}{3\sqrt{n}}\right)\right]^{-1}, \tag{2.7}$$

and

$$\tilde{\alpha}(n,\lambda) \geq \left[\rho + \gamma\left(\frac{\Phi(\alpha)}{\phi(\alpha)} + \frac{2}{3\sqrt{n} + \frac{1}{\phi(\alpha)}\frac{1}{12n-1}}\right)\right]^{-1}. \tag{2.8}$$

Using equation (2.5) we can express the continuous Erlang-C formula and its bounds in terms of β and λ. Define

$$\tilde{\alpha}_\beta(\beta,\lambda) = \tilde{\alpha}(\lambda + \beta\sqrt{\lambda}, \lambda)$$

as the continuous Erlang-C formula with respect to β, where $\tilde{\alpha}(\cdot,\cdot)$ is defined in equation (2.1). Let $UB(\beta,\lambda)$ represent the upper bound on the Erlang-C, as given on the right-hand side of inequality (2.7), and let $LB(\beta,\lambda)$ represent the lower bound, as given on the right-hand side of inequality (2.8).

We now present a series of lemmas which we use to prove our main results on asymptotic approximations for our staffing problems of interest.

Lemma 3. The upper bound for the Erlang-C formula, $UB(\beta,\lambda)$, as defined on the right-hand side of inequality (2.7), is strictly decreasing in λ for any $\beta > 0$.

Proof. To show $UB(\beta,\lambda)$ is strictly decreasing in λ, it suffices to show the denominator of (2.7), $\rho + \gamma\left(\frac{\Phi(\alpha)}{\phi(\alpha)} + \frac{2}{3\sqrt{n}}\right)$, is strictly increasing in λ for any $\beta > 0$. First note that

$$\rho + \gamma\left(\frac{2}{3\sqrt{n}}\right) = \frac{3\lambda + 2\beta\sqrt{\lambda}}{3\left(\lambda + \beta\sqrt{\lambda}\right)}$$

is strictly increasing in λ for any $\beta > 0$, since

$$\frac{\partial\left[\frac{3\lambda+2\beta\sqrt{\lambda}}{3(\lambda+\beta\sqrt{\lambda})}\right]}{\partial\lambda} = \frac{\beta\sqrt{\lambda}}{6(\lambda+\beta\sqrt{\lambda})^2} > 0 \quad \forall\beta,\lambda > 0.$$

Thus, we only need to prove that $\frac{\gamma\Phi(\alpha)}{\phi(\alpha)}$ is non-decreasing in λ for any $\beta > 0, \lambda > 0$. Let

$$f(\beta,\lambda) = \gamma\frac{\Phi(\alpha)}{\phi(\alpha)} = \beta\sqrt{\frac{\lambda}{\lambda+\beta\sqrt{\lambda}}}\frac{\Phi(\alpha)}{\phi(\alpha)}.$$

We can show that $f(\beta,\lambda)$ is strictly increasing by showing that $\frac{\partial f}{\partial \lambda} > 0$ for any $\beta > 0, \lambda > 0$. We have

$$\frac{\partial f}{\partial \lambda} = \frac{\beta^2}{4(\lambda+\beta\sqrt{\lambda})\sqrt{\lambda+\beta\sqrt{\lambda}}}\frac{\Phi(\alpha)}{\phi(\alpha)} + \beta\sqrt{\frac{\lambda}{\lambda+\beta\sqrt{\lambda}}}\frac{[\phi(\alpha)+\Phi(\alpha)\alpha]\frac{\partial\alpha}{\partial\lambda}}{\phi(\alpha)}. \quad (2.9)$$

From (2.9), it is obvious that $\frac{\partial f}{\partial \lambda} > 0$ for any $\beta > 0, \lambda > 0$, if $\frac{\partial \alpha}{\partial \lambda} > 0$ for any $\beta > 0, \lambda > 0$. We have

$$\frac{\partial \alpha}{\partial \lambda} = \frac{-\frac{\beta}{\sqrt{\lambda}} + \left(1+\frac{\beta}{2\sqrt{\lambda}}\right)\ln\left(1+\frac{\beta}{\sqrt{\lambda}}\right)}{\sqrt{-2\left(\beta\sqrt{\lambda}+\lambda\right)\left(1-\frac{\lambda}{\beta\sqrt{\lambda}+\lambda}+\ln\left(\frac{\lambda}{\beta\sqrt{\lambda}+\lambda}\right)\right)}}. \quad (2.10)$$

The denominator of (2.10) is greater than 0, so it suffices to show

$$g(\beta,\lambda) = -\frac{\beta}{\sqrt{\lambda}} + \left(1+\frac{\beta}{2\sqrt{\lambda}}\right)\ln\left(1+\frac{\beta}{\sqrt{\lambda}}\right) > 0, \quad \forall \beta > 0, \lambda > 0.$$

We have

$$\lim_{\lambda\to\infty} g(\beta,\lambda) = 0, \quad \forall \beta > 0,$$

and

$$\frac{\partial g}{\partial \lambda} = \frac{\beta}{2\lambda\sqrt{\lambda}}\left[\frac{\frac{\beta}{\sqrt{\lambda}}}{1+\frac{\beta}{\sqrt{\lambda}}} - \ln\left(1+\frac{\beta}{\sqrt{\lambda}}\right)\right].$$

Since

$$\frac{x}{1+x} < \ln(1+x), \forall x > 0,$$

we have

$$\frac{\frac{\beta}{\sqrt{\lambda}}}{1+\frac{\beta}{\sqrt{\lambda}}} - \ln\left(1+\frac{\beta}{\sqrt{\lambda}}\right) < 0, \quad \forall \beta > 0, \lambda > 0.$$

Thus we have $\frac{\partial g}{\partial \lambda} < 0$ and

$$\lim_{\lambda\to\infty} g\left(\beta,\lambda\right) = 0, \quad \forall \beta > 0.$$

This proves $g(\beta,\lambda) > 0, \forall \beta > 0, \lambda > 0$. □

Lemma 4. Let $n = \lambda + \beta\sqrt{\lambda}$ and let γ be defined as in (2.6). Then $\gamma/(12n-1)$ converges uniformly to 0, in β, as $\lambda \to \infty$. That is

$$\lim_{\lambda\to\infty} \sup_{\beta>0} \frac{\gamma}{12n-1} = 0.$$

Proof. With

$$g(\beta,\lambda) = \frac{\gamma}{(12n-1)} = \frac{\beta\sqrt{\lambda/\left(\lambda+\beta\sqrt{\lambda}\right)}}{12\left(\lambda+\beta\sqrt{\lambda}\right)-1},$$

we have that $g(\beta,\lambda) > 0$ and continuous in (β,λ) for all $\beta > 0, \lambda > 10$. (Below, we let λ grow large. The value "10" here simply serves as a sufficiently large lower bound we use in establishing the desired result.) We have

$$\frac{\partial g(\beta,\lambda)}{\partial\beta} = -\frac{12\beta\sqrt{\lambda}\sqrt{\frac{\lambda}{\lambda+\beta\sqrt{\lambda}}}}{\left(-1+12\left(\lambda+\beta\sqrt{\lambda}\right)\right)^2} + \frac{\sqrt{\frac{\lambda}{\lambda+\beta\sqrt{\lambda}}}}{-1+12\left(\lambda+\beta\sqrt{\lambda}\right)} - \frac{\beta\lambda^{3/2}}{2\sqrt{\frac{\lambda}{\lambda+\beta\sqrt{\lambda}}}\left(\lambda+\beta\sqrt{\lambda}\right)^2\left(-1+12\left(\lambda+\beta\sqrt{\lambda}\right)\right)}.$$

Forming

$$\frac{\partial g(\beta,\lambda)}{\partial\beta} = 0,$$

yields, after some algebra,

$$\frac{\sqrt{\frac{\sqrt{\lambda}}{\beta+\sqrt{\lambda}}}\left(-12\beta^2\sqrt{\lambda}+\beta(-1+12\lambda)+2\sqrt{\lambda(-1+12\lambda)}\right)}{-1+12\beta\sqrt{\lambda}+12\lambda} = 0,$$

or equivalently,

$$-12\beta^2\sqrt{\lambda}+\beta(-1+12\lambda)+2\sqrt{\lambda(-1+12\lambda)} = 0.$$

Then we have

$$\widehat{\beta} = \frac{-1 + 12\lambda + \sqrt{1 - 120\lambda + 1296\lambda^2}}{24\sqrt{\lambda}},$$

as the only positive root of this equation. Also we have $g(0, \lambda) = 0, \forall \lambda > 10$,

$$\lim_{\beta \to \infty} g(\beta, \lambda) = 0, \quad \forall \lambda > 10,$$

and $g(\widehat{\beta}, \lambda) > 0, \forall \lambda > 10$. Thus $\widehat{\beta}$ is the global maximizer of $g(\beta, \lambda)$ for any $\lambda > 10$. That is,

$$g(\widehat{\beta}, \lambda) = \max_{\beta > 0} g(\beta, \lambda), \quad \forall \lambda > 10.$$

Since

$$\lim_{\lambda \to \infty} g(\widehat{\beta}, \lambda) = 0,$$

we have

$$\lim_{\lambda \to \infty} \sup_{\beta > 0} g(\beta, \lambda) = 0.$$

This completes the proof. □

Lemma 5. Let $n = \lambda + \beta\sqrt{\lambda}$ and let $\rho, \gamma, \alpha, \phi(\cdot)$ and $\Phi(\cdot)$ be defined as in Theorem 2. Then $\rho\phi(\alpha) + \gamma\Phi(\alpha)$ is strictly increasing in β for any sufficiently large λ.

Proof. To show $\rho\phi(\alpha) + \gamma\Phi(\alpha)$ is strictly increasing in β for any sufficiently large λ, it suffices to show $\rho\phi(\alpha) + \gamma\Phi(\alpha)$ is strictly increasing in n for any sufficiently large λ, since β and n satisfy a linear relationship with a positive slope. Let $h(n, \lambda) = \rho\phi(\alpha) + \gamma\Phi(\alpha)$. We have

$$\frac{\partial h(n, \lambda)}{\partial n} = \frac{-\lambda\phi(\alpha)}{n^2} + \Phi(\alpha)\frac{n + \lambda}{2n\sqrt{n}} + \frac{\lambda\phi(\alpha)\ln\left(\frac{\lambda}{n}\right)}{n} - \frac{(n - \lambda)\phi(\alpha)\ln\left(\frac{\lambda}{n}\right)}{\alpha\sqrt{n}}.$$

First note that

$$\frac{-\lambda\phi(\alpha)}{n^2} + \Phi(\alpha)\frac{n+\lambda}{2n\sqrt{n}} > \frac{-\lambda\phi(\alpha) + n\Phi(\alpha)}{n^2},$$

since $n > \lambda \geq 1$. (Here, we take "1" as a lower bound on λ since we establish a result for λ that is sufficiently large.) Also since

$$\frac{-\lambda\phi(\alpha) + n\Phi(\alpha)}{n^2} \geq \frac{-\lambda\phi(0) + n\Phi(0)}{n^2} > 0,$$

we have

$$\frac{-\lambda\phi(\alpha)}{n^2} + \Phi(\alpha)\frac{n+\lambda}{2n\sqrt{n}} > 0.$$

It remains then to show that

$$\frac{\lambda\phi(\alpha)\ln\left(\frac{\lambda}{n}\right)}{n} - \frac{(n-\lambda)\phi(\alpha)\ln\left(\frac{\lambda}{n}\right)}{\alpha\sqrt{n}} \geq 0 \quad \forall\beta > 0, \lambda > 0.$$

Some algebra demonstrates that it is equivalent to show $\lambda\alpha \leq \sqrt{n}(n-\lambda)$. As shown in [25],

$$\alpha = \beta - \frac{1}{6}\beta^2\frac{1}{\sqrt{\lambda}} + O\left(1/\lambda\right).$$

So we have $0 < \alpha < \beta$ for sufficiently large λ. Thus

$$\lambda\alpha < \lambda\beta = \sqrt{\lambda}(n-\lambda) < \sqrt{n}(n-\lambda),$$

since $\lambda < n$. □

Lemma 6. Let $n = \lambda + \beta\sqrt{\lambda}$ and let $\rho, \gamma, \alpha, \phi(\cdot)$ and $\Phi(\cdot)$ be defined as in Theorem 2. Then $\rho\phi(\alpha) + \gamma\Phi(\alpha) + \frac{2\gamma\phi(\alpha)}{3\sqrt{n}} + \frac{\gamma}{(12n-1)}$ is uniformly bounded away from 0 for all sufficiently large λ, and all $\beta > 0$. Specifically,

$$\inf_{\lambda \geq M, \beta > 0} \rho\phi(\alpha) + \gamma\Phi(\alpha) + \frac{2\gamma\phi(\alpha)}{3\sqrt{n}} + \frac{\gamma}{(12n-1)} > 0, \tag{2.11}$$

where M is a sufficiently large value.

Proof. All four terms in the formula on the left-hand side of inequality (2.11) are non-negative for all $\beta > 0$, $\lambda \geq 1$. Thus it suffices to show that $h(\beta, \lambda) = \rho\phi(\alpha)+\gamma\Phi(\alpha)$ is uniformly bounded away from 0 for all sufficiently large λ and $\beta > 0$, since $\phi(\cdot)$ is positive and bounded. From Lemma 5, we know $h(\beta, \lambda)$ is strictly increasing in β for all sufficiently large λ and $h(0, \lambda) = \phi(0) > 0$. Thus we have that $h(\beta, \lambda)$ is uniformly bounded away from 0 for all sufficiently large λ and all $\beta > 0$. □

Lemma 7. Let $\lambda > 0$ and let $UB(\beta, \lambda)$ be defined by the formula on the right-hand side of inequality (2.7). Then $UB(\beta, \lambda)$ is strictly decreasing in β for any $\lambda, \beta > 0$.

Proof. Let $UB_n(n, \lambda) = UB(\frac{n-\lambda}{\sqrt{\lambda}}, \lambda)$. Since $n = \lambda + \beta\sqrt{\lambda}$, to show $UB(\beta, \lambda)$ is strictly decreasing in β for any $\lambda > 0$, it suffices to show $UB_n(n, \lambda)$ is strictly decreasing in n. So it suffices to show

$$UB_n^{-1}(n, \lambda) = \frac{\lambda}{n} + \frac{n-\lambda}{\sqrt{n}}\left(\frac{\Phi(\alpha)}{\phi(\alpha)} + \frac{2}{3\sqrt{n}}\right),$$

is strictly increasing in n for any λ, where α is given in (2.4). Now,

$$\frac{\partial UB_n^{-1}(n, \lambda)}{\partial n} = \frac{-\lambda}{3n^2} + \left(\frac{\lambda+n}{2n\sqrt{n}}\right)\left(\frac{\Phi(\alpha)}{\phi(\alpha)}\right) + \left(\frac{n-\lambda}{\sqrt{n}}\right)\left(\frac{\phi(\alpha)+\Phi(\alpha)\alpha}{\phi(\alpha)}\right)\left(\frac{\partial\alpha}{\partial n}\right). \quad (2.12)$$

Here,

$$\frac{\partial\alpha}{\partial n} = \frac{-\ln\frac{\lambda}{n}}{\alpha}.$$

Since $n > \lambda$, we have $\frac{\partial\alpha}{\partial n} > 0$ and so the third term in (2.12) is non-negative. Since $\Phi(\alpha)/\phi(\alpha)$ is strictly increasing in α and α is strictly increasing in n, we

have that $\Phi(\alpha)/\phi(\alpha)$ is strictly increasing in n. Thus the first two terms are greater than

$$\frac{-\lambda}{3n^2} + \frac{\lambda + n}{2n\sqrt{n}} \frac{\Phi(0)}{\phi(0)} > 0,$$

which completes the proof. □

Theorem 8. The upper bound of (2.7) and lower bound of (2.8) on the Erlang-C formula uniformly converge to the Erlang-C formula as $\lambda \to \infty$. That is

$$\lim_{\lambda\to\infty} \sup_{\beta>0} [\tilde{\alpha}_\beta(\beta, \lambda) - LB(\beta, \lambda)] = 0$$

and

$$\lim_{\lambda\to\infty} \sup_{\beta>0} [UB(\beta, \lambda) - \tilde{\alpha}_\beta(\beta, \lambda)] = 0.$$

Proof. To prove the uniform convergence of the bounds to the Erlang-C formula, we only need to show the upper bound, $UB(\beta, \lambda)$, uniformly converges to the lower bound, $LB(\beta, \lambda)$, i.e.,

$$\lim_{\lambda\to\infty} \sup_{\beta>0} [UB(\beta, \lambda) - LB(\beta, \lambda)] = 0.$$

With $n = \lambda + \beta\sqrt{\lambda}$, we have

$$\begin{aligned} UB(\beta, \lambda) - LB(\beta, \lambda) &= \frac{1}{\rho + \gamma\left(\frac{\Phi(\alpha)}{\phi(\alpha)} + \frac{2}{3\sqrt{n}}\right)} - \frac{1}{\rho + \gamma\left(\frac{\Phi(\alpha)}{\phi(\alpha)} + \frac{2}{3\sqrt{n}}\right) + \frac{\gamma}{\phi(\alpha)(12n-1)}} \\ &= \frac{1}{\rho + \gamma\left(\frac{\Phi(\alpha)}{\phi(\alpha)} + \frac{2}{3\sqrt{n}}\right)} \cdot \frac{\frac{\gamma}{\phi(\alpha)(12n-1)}}{\rho + \frac{\gamma\Phi(\alpha)}{\phi(\alpha)} + \frac{2\gamma}{3\sqrt{n}} + \frac{\gamma}{\phi(\alpha)(12n-1)}}. \end{aligned}$$

From Lemma 7, we have that

$$\frac{1}{\rho + \gamma\left(\frac{\Phi(\alpha)}{\phi(\alpha)} + \frac{2}{3\sqrt{n}}\right)}$$

is strictly decreasing in β for any fixed $\lambda > 0$ and

$$\left[\rho + \gamma\left(\frac{\Phi(\alpha)}{\phi(\alpha)} + \frac{2}{3\sqrt{n}}\right)\right]^{-1}\Bigg|_{\beta=0} = 1 \quad \forall \lambda > 0.$$

Thus to show

$$\lim_{\lambda\to\infty} \sup_{\beta>0} UB(\beta,\lambda) - LB(\beta,\lambda) = 0,$$

it suffices to prove

$$\lim_{\lambda\to\infty} \sup_{\beta>0} \frac{\gamma}{(12n-1)} = 0,$$

and

$$\inf_{\lambda\geq 1,\beta>0} \phi(\alpha)\left(\rho + \frac{\gamma\Phi(\alpha)}{\phi(\alpha)} + \frac{2\gamma}{3\sqrt{n}} + \frac{\gamma}{\phi(\alpha)(12n-1)}\right) > 0,$$

which hold via Lemma 4 and Lemma 6. □

Theorem 8 plays a very important role in proving the asymptotic optimality property of our approximate optimal staffing decision in Chapter 3 and Chapter 4.

2.2 $M/M/n + M$ Queue

The $M/M/n$ queue incorporates busy signals, but does not allow a customer to abandon the queue before receiving service. The model that incorporates both busy signals and abandonment is the so-called $M/M/n/k+G$ queue. Like the $M/M/n$ queue, the $M/M/n/k+G$ system models a call center with n servers, and it also assumes a Poisson arrival process and exponentially distributed service times. The service times are independent of each other and of the arrival process. A customer entering the system begins receiving

service immediately if a server is idle, or, if all servers are busy, the customer waits in the queue for an agent to become free. We again assume the service discipline to be FCFS. In the $M/M/n/k+G$ model, a customer may abandon the system after waiting in the system for some time. The amount of time that a customer will wait for service is called patience. The "$+G$" notation indicates that patience has a general distribution, is independent and identically distributed (i.i.d.) over customers and is independent of other random elements. The "k"in the notation indicates the buffer size; that is, no more than k customers can wait in queue. A customer arriving when the system already has k customers in queue does not enter the system. The simplest $M/M/n/k+G$ queueing system is the $M/M/n+M$ system, or the so-called Erlang-A system. The Erlang-A system assumes an infinite buffer size and an exponentially distributed patience distribution. The Erlang-A queueing model is shown in Figure 2.2, where λ stands for arrival rate, μ stands for service rate, θ stands for abandonment rate and n stands for the number of service agents. The mathematical details on the stationary distribution governing the number of customers in the queue of the Erlang-A system can be found in Palm [29], Palm [30], Riordan [32] and Mandelbaum and Zeltyn [28]. The following gives the formulas for the steady-state metrics including the probability an arriving customer waits, and the probability a customer abandons the system, which can be found in the work of Mandelbaum and Zeltyn [28]:

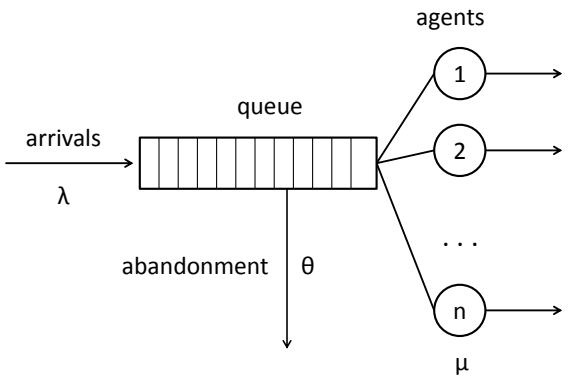

Figure 2.2: Erlang-A Model

Let

$$A(x,y) = \frac{xe^y}{y^x}\int_0^y t^{x-1}e^{-t}dt = 1 + \sum_{j=1}^{\infty}\frac{y^j}{\prod_{k=1}^{j}(x+k)}, \quad x > 0, y \geq 0,$$

and

$$E_{1,n} = \frac{\frac{(\lambda/\mu)^n}{n!}}{\sum_{j=0}^{n}\frac{(\lambda/\mu)^j}{j!}}. \tag{2.13}$$

Then, the probability an arriving customer waits is

$$\mathbb{P}\{Wait > 0\} = \frac{A\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta}\right) \cdot E_{1,n}}{1 + \left(A\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta}\right) - 1\right) \cdot E_{1,n}};$$

and the probability of abandonment given that all servers are busy is

$$\mathbb{P}\{Abandonment | Wait > 0\} = \frac{1}{\rho A\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta}\right)} + 1 - \frac{1}{\rho},$$

where $\rho = \frac{\lambda}{n\mu}$.

The probability that a customer abandons the system, $\mathbb{P}\{Abandonment\}$, is given by

$$\mathbb{P}\{Wait > 0\} \times \mathbb{P}\{Abandonment | Wait > 0\},$$

which is the so-called Erlang-A formula.

We now introduce a continuous extension of the Erlang-A formula. This continuous Erlang-A formula, which allows the number of servers n to take a non-integer value, gives the same value as the Erlang-A formula for the integer points of n, and is defined in the same way of Erlang-A formula, only that in the definition of the continuous version of the Erlang-A formula, we replace the formula (2.13) with its continuous extension:

$$\widetilde{E_{1,n}} = \frac{\frac{(\lambda/\mu)^n}{\Gamma(n+1)}}{\frac{e^{\lambda/\mu}\Gamma(n+1,\lambda/\mu)}{\Gamma(n+1)}} = \frac{(\lambda/\mu)^n}{e^{\lambda/\mu}\Gamma(n+1,\lambda/\mu)}. \tag{2.14}$$

With (2.14), the continuous version of the Erlang-A formula is given as:

$$\widetilde{Erlang}A(n;\lambda,\mu,\theta) = \frac{A\left(\frac{n\mu}{\theta},\frac{\lambda}{\theta}\right)\cdot\widetilde{E_{1,n}}}{1+\left(A\left(\frac{n\mu}{\theta},\frac{\lambda}{\theta}\right)-1\right)\cdot\widetilde{E_{1,n}}} \times \left(\frac{1}{\rho A\left(\frac{n\mu}{\theta},\frac{\lambda}{\theta}\right)}+1-\frac{1}{\rho}\right). \tag{2.15}$$

The continuous version of the Erlang-A formula allows us to consider the staffing problem with continuous decision variables, which facilitates some of our analysis. Later we use as a metric the probability that a customer abandons the system, and in our work, we will use the continuous version of the Erlang-A formula. Also, for simplicity, we will assume $\mu = 1$ in our work, when we apply the Erlang-A formula.

Chapter 3

Deterministic Arrival-rate Problems

In this chapter, we use approximations discussed in Chapter 2 in order to approximate service quality in an analysis of the trade-off between staffing cost and service quality for the known arrival rate situation. Our call center manager has two competing concerns. First the manager is concerned with the staffing cost, and hence would tend to hire as few servers as possible. Second, the manager is concerned with service quality, which will be poor if an insufficient number of servers are hired. In this chapter, we start by building a model for a single-class single-station service system, and then extend the model to handle a multi-class multi-station system. We prove asymptotic optimality of solutions for our approximate model for both the single-class single-station system and the multi-class multi-station system, where our asymptotic analysis has the arrival rate grow large.

3.1 Queueing Systems with Deterministic Arrival Rates

We consider a call center that has either a single $M/M/n$ queue representing a single service station, as depicted in Figure 3.1, or L parallel $M/M/n$ queues representing a multi-station system, as depicted in Figure 3.2. In this

chapter we use the probability that a customer must wait to receive service to measure the quality of service. We model the trade-off between the staffing cost and the probability of waiting for deterministic arrival-rate systems and give our asymptotic optimality results.

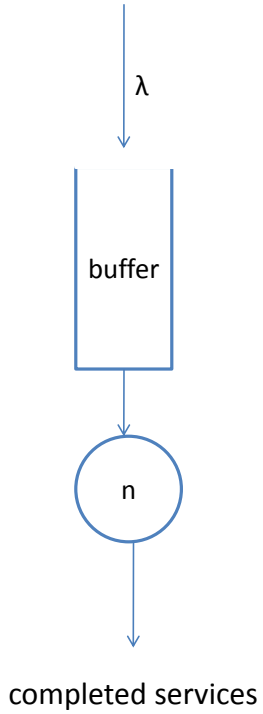

Figure 3.1: Single-Queue System - Deterministic Arrival Rate

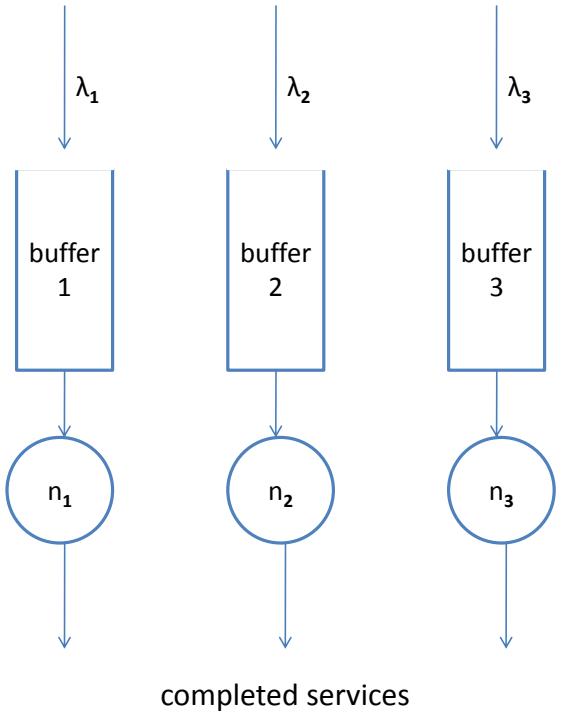

Figure 3.2: Multi-Queue System - Deterministic Arrival Rates

3.1.1 Single-station System

We begin with the single-station system. Because of the manager's competing measures, we face a bi-criteria optimization problem, in which we want to simultaneously minimize the staffing cost and the probability of inducing customer waiting. Let $\bar{c}(n)$ be the staffing cost function, and assume $\bar{c}(n)$ is strictly increasing in the staffing level n. We assume that our $M/M/n$ queue is in steady state and we let $\overline{wait}(n, \lambda)$ denote the number of customers waiting in the queue for service when there are n servers and the arrival rate is λ. Our bi-criteria model for this call center problem is:

$$\text{vmin}_{n \in \mathbb{Z}_+} \quad \left[\bar{c}(n), \mathbb{P}\left\{\overline{wait}(n, \lambda) > 0\right\}\right], \tag{3.1}$$

where "vmin" denotes vector minimization and a solution of model (3.1) corresponds to the family of staffing levels that falls on the efficient frontier, and $\mathbb{Z}_+$ denotes the set of non-negtive integers.

More generally, we are interested in obtaining asymptotic results. We consider a sequence of problems of the form (3.1) with $\lambda \to \infty$. As λ goes to ∞, the staffing level, n, also goes to ∞, and so does the staffing cost $c(n)$. Hence, we need to reformulate (3.1) to obtain a well-posed model. In the heavy traffic regime, we use the square-root staffing policy that we discuss in Chapter 2 and that is given in Halfin and Whitt [22]. We use $\lambda + \beta\sqrt{\lambda}$ to replace n in (3.1), and rewrite the model in β:

$$\text{vmin}_{\beta \geq 0} \quad \left[c(\beta), \mathbb{P}\left\{wait(\beta, \lambda) > 0\right\}\right], \tag{3.2}$$

where $c(\beta)$ is the cost function parameterized in β rather than in n, the number of customers, and $wait(\beta, \lambda)$ denotes the random number of customers waiting in the queue for service parameterized in β and λ. Our motivation for this is as follows, we know that in order to have a well-behaved system as λ grows large we must staff according to the square-root staffing rule. (See Theorem 1 in Chapter 2.) Hence, the critical question in this setting is the level of the safety parameter, β, rather than the absolute number of servers, which will grow large with λ. We assume $c(\beta)$ is strictly increasing in its argument, and hence for any fixed value of λ, models (3.1) and (3.2) are equivalent. (Recall, our goal in these bi-criteria models is to form the efficient frontier of solutions.) Moreover, the optimal value of β does not grow large as λ grows large, and hence the asymptotics associated with model (3.2) have finite limits.

One way to solve model (3.2) is to make one component of the objective function a constraint. For example, we can solve the bi-criteria problem by solving a parameterized family of models:

$$\min_{\beta \geq 0} \quad c(\beta) \qquad \text{s.t.} \quad \mathbb{P}\left\{wait(\beta, \lambda) > 0\right\} \leq \epsilon, \tag{3.3}$$

parameterized in the risk level threshold, ϵ, where $0 < \epsilon < 1$.

For each ϵ, by solving model (3.3), we obtain an optimal β. The staffing cost and the probability of waiting corresponding to the optimal β give one point on the efficient frontier of the bi-criteria problem. By varying ϵ from 0 to 1, we obtain all the points on the efficient frontier.

Another way to solve model (3.2) is to use a weighted objective function

approach. That is we solve the bi-criteria problem by solving a parameterized family of models:

$$\min_{\beta \geq 0} \quad c(\beta) + \delta \mathbb{P}\left\{wait(\beta, \lambda) > 0\right\}, \tag{3.4}$$

parameterized by $\delta > 0$, the weight on the second term in the objective function. Solving model (3.4), by varying δ, we obtain all the extreme points of the convex hull of the efficient frontier [31].

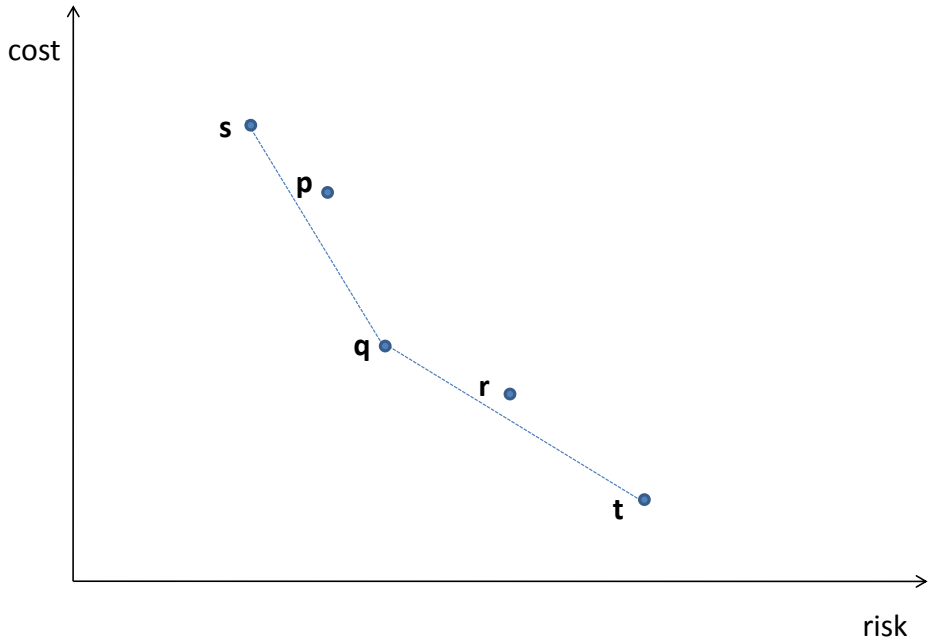

Figure 3.3: Efficient Frontier

We use the example given in Figure 3.3 to explain the relationship between model (3.3) and model (3.4). In Figure 3.3, we assume the points p, q, r, s and t correspond to Pareto efficient solutions to the bi-criteria problem under consideration. If we use model (3.3) to solve the bi-criteria problem, then by varying ϵ, we achieve all solutions on the efficient frontier. So, we achieve all five points, p, q, r, s and t. On the other hand, if we use model (3.4), unless the efficient frontier is convex, we will not achieve all five of these points. However, the solutions which are extreme points of the efficient frontier will be achieved. That is, points s, q and t will be achieved by solving model (3.4) and varying δ.

Model (3.3) directly describes what is typically viewed as the practical need. Generally, call center managers try to find a staffing level that minimizes the staffing cost while maintaining a certain service level. However, in our view there is insight to be gained by forming the efficient frontier to better understand cost-quality tradeoffs. This is particularly true when contractual service levels have not yet been determined. In what follows we use either model (3.3) or (3.4) to present results, depending on which is more convenient.

For the $M/M/n$ queue, the probability of waiting is given by the continuous Erlang-C formula, that is,

$$\mathbb{P}\left\{wait(\beta,\lambda) > 0\right\} = \tilde{\alpha}_\beta(\beta,\lambda).$$

Here, $\tilde{\alpha}_\beta(\beta,\lambda) = \tilde{\alpha}(\lambda + \beta\sqrt{\lambda},\lambda)$ and $\tilde{\alpha}(\cdot,\cdot)$ is defined in equation (2.1).

Using the continuous Erlang-C formula, for any fixed λ, define model

F_λ as:

$$\min_{\beta \geq 0} \quad c(\beta) \qquad \text{s.t. } \tilde{\alpha}_\beta(\beta, \lambda) \leq \epsilon. \tag{3.5}$$

As λ varies, we obtain a sequence of models $\{F_\lambda\}$. The Erlang-C formula can be hard to handle numerically, especially when the arrival rate grows large. So we build an approximate model by replacing the Erlang-C formula with the upper bound, $UB(\beta, \lambda)$, defined by the equation on the right-hand side of (2.7), except that n is replaced by $\lambda + \beta\sqrt{\lambda}$. Using $UB(\beta, \lambda)$ we define our approximate model G_λ as:

$$\min_{\beta \geq 0} \quad c(\beta) \qquad \text{s.t. } UB(\beta, \lambda) \leq \epsilon. \tag{3.6}$$

Notice that any feasible solution of model G_λ is also feasible for model F_λ.

Theorem 10 below characterizes the relationship between optimal solutions to models G_λ and F_λ and further characterizes the limit of these solutions as λ grows large. Before turning to Theorem 10, we first provide a supporting lemma.

Lemma 9. Let $\lambda > 0$ and $\tilde{\alpha}(n, \lambda)$ be the continuous Erlang-C formula as defined in (2.1). Define $\tilde{\alpha}_\beta(\beta, \lambda) = \tilde{\alpha}(\lambda + \beta\sqrt{\lambda}, \lambda)$. Then $\tilde{\alpha}_\beta(\beta, \lambda)$ is strictly decreasing in β for any $\lambda > 0$ and any $\beta > 0$ that satisfy $\lambda + \beta\sqrt{\lambda} \geq 1$.

Proof. To prove $\tilde{\alpha}_\beta(\beta, \lambda)$ is strictly decreasing in β, it suffices to show that $\tilde{\alpha}(n, \lambda)$ is strictly decreasing in n. Jagers and Van Doorn [24] prove that the continuous Erlang-C formula, $\tilde{\alpha}(n, \lambda)$, is convex in n. Also, we know that

$\tilde{\alpha}(n, \lambda)$ is strictly decreasing in n on the positive integers. This implies $\tilde{\alpha}(n, \lambda)$ is strictly decreasing in n for all $n > 1$, otherwise its epigraph is not convex. $\square$

Theorem 10. Let $\lambda > 0$. Let the optimal solution of F_λ, as defined in (3.5), be β_λ^F and let the optimal solution of G_λ, as defined in (3.6), be β_λ^G. Then $\beta_\lambda^G \geq \beta_\lambda^F$, $\forall \lambda > 0$, and there exists a finite β^* such that

$$\lim_{\lambda \to \infty} \beta_\lambda^G = \lim_{\lambda \to \infty} \beta_\lambda^F = \beta^*.$$

Proof. The objective function, $c(\beta)$, is strictly increasing in β and by Lemma 7 and Lemma 9, $\tilde{\alpha}_\beta(\beta, \lambda)$ and $UB(\beta, \lambda)$ are strictly decreasing and continuous in β. Hence, the unique optimal solution of models (3.5) and (3.6) are defined by requiring the respective constraints to hold with equality. That is, β_λ^F solves $\tilde{\alpha}_\beta(\beta, \lambda) = \epsilon$ and β_λ^G solves $UB(\beta, \lambda) = \epsilon$. We know that $UB(\beta, \lambda)$ and $\tilde{\alpha}_\beta(\beta, \lambda)$ have range (0,1] and we have

$$UB(0, \lambda) = \tilde{\alpha}_\beta(0, \lambda) = 1$$

and

$$\lim_{\beta \to \infty} UB(\beta, \lambda) = \lim_{\beta \to \infty} \tilde{\alpha}_\beta(\beta, \lambda) = 0.$$

Also $UB(\beta, \lambda)$ and $\tilde{\alpha}_\beta(\beta, \lambda)$ are continuous and strictly decreasing in β on $[0, \infty)$. So, the optimal solutions β_λ^F and β_λ^G exist and are unique for any $\epsilon > 0$ and $\lambda > 0$. From Lemma 3, we have $\beta_{\lambda_1}^G > \beta_{\lambda_2}^G \geq 0$, for any $\lambda_2 > \lambda_1$. This indicates $\lim_{\lambda \to \infty} \beta_\lambda^G$ exists and is finite. Let $\lim_{\lambda \to \infty} \beta_\lambda^G = \beta^*$. Also, since $UB(\beta, \lambda) \geq \tilde{\alpha}_\beta(\beta, \lambda)$ for any $\beta > 0, \lambda > 0$, we have $\beta_\lambda^G \geq \beta_\lambda^F$ for any $\lambda > 0$. This together with the fact that $\{\beta_\lambda^G\}$ is a bounded sequence, indicate that

$\{\beta_\lambda^F\}$ is a bounded sequence. So $\{\beta_\lambda^F\}$ has at least one subsequence that has a finite limit. For any subsequence $\{\beta_{\lambda'}^F\}$ with a limit and its corresponding limit $\hat{\beta}$, we have

$$\lim_{\lambda' \to \infty} \tilde{\alpha}_\beta(\hat{\beta}, \lambda') = \epsilon.$$

Also, for any β, we have

$$\lim_{\lambda \to \infty} (\tilde{\alpha}_\beta(\beta, \lambda) - UB(\beta, \lambda)) = 0.$$

This indicates that

$$\lim_{\lambda' \to \infty} UB(\hat{\beta}, \lambda') = \lim_{\lambda' \to \infty} \tilde{\alpha}_\beta(\hat{\beta}, \lambda') = \epsilon.$$

Since

$$\lim_{\lambda' \to \infty} UB(\beta^*, \lambda') = \epsilon,$$

and there is a unique β satisfies

$$\lim_{\lambda' \to \infty} UB(\beta, \lambda') = \epsilon,$$

this indicates that $\hat{\beta} = \beta^*$. This indicates that all subsequences of $\{\beta_\lambda^F\}$ have the same limit point, β^*. Thus the limit of $\{\beta_\lambda^F\}$ exists and is β^*. That is

$$\lim_{\lambda \to \infty} \beta_\lambda^G = \lim_{\lambda \to \infty} \beta_\lambda^F = \beta^*.$$

□

3.1.2 Multi-station System

We now extend our development to a multi-station system. We suppose that we have L $M/M/n$ queues in parallel, each with its own arrival process

with rate λ_i, $i = 1, \ldots, L$, and each with n_i servers, determined by β_i, $i = 1, \ldots, L$, via the square-root staffing rule. Then we formulate:

$$\min_{\beta \geq 0} \quad \sum_{i=1}^{L} c_i(\beta_i) + \delta \mathbb{P}\left\{ \bigcup_{i=1}^{L} \{wait_i(\beta_i, \lambda_i) > 0\} \right\}, \tag{3.7}$$

where, $\beta = (\beta_1, \ldots, \beta_L)$ and $wait_i(\beta_i, \lambda_i)$ represents the random number of customers waiting in queue i for service, analogous to the definition of $wait(\beta, \lambda)$ in model (3.3).

We assume the L queues operate independently. Then the above model is equivalent to:

$$\min_{\beta \geq 0} \quad \sum_{i=1}^{L} c_i(\beta_i) + \delta \left(1 - \prod_{i=1}^{L} \left(1 - \mathbb{P}\left\{ wait_i\left(\beta_i, \lambda_i\right) > 0 \right\} \right) \right). \tag{3.8}$$

As in the single-station system, under the $M/M/n$ assumption, we can formulate an equivalent model where the probability of waiting is calculated by the continuous Erlang-C formula, and we again denote this by model F_λ:

$$\min_{\beta \geq 0} \quad \sum_{i=1}^{L} c_i(\beta_i) + \delta \left(1 - \prod_{i=1}^{L} \left(1 - \tilde{\alpha}_\beta\left(\beta_i, \lambda_i\right) \right) \right). \tag{3.9}$$

Again following an analogous development to our single-station system, we build an approximate model for (3.9) by using $UB(\beta, \lambda)$. The approximate model G_λ is:

$$\min_{\beta \geq 0} \quad \sum_{i=1}^{L} c_i(\beta_i) + \delta \left(1 - \prod_{i=1}^{L} \left(1 - UB\left(\beta_i, \lambda_i\right) \right) \right). \tag{3.10}$$

Denote the objective function of model F_λ as $f_\lambda(\cdot)$, and the optimal solution of F_λ as the L-vector β_λ^F. Similarly, denote the objective function

of model G_λ as $g_\lambda(\cdot)$, and the optimal solution of G_λ as β^G_λ. The following lemma and theorem characterize asymptotic optimality of the multi-station system as the arrival rate vector grows large. We let the arrival rates grow in the following way: We assume there are initial values of arrival rates for all queues. Let the initial vector of rates be $\lambda^0 = (\lambda^0_1, \ldots, \lambda^0_L)$. Indexing the sequence of systems under consideration with positive integers, we assume the arrival rate for the m^{th} system is $\lambda^m = m\lambda^0$. Then as $m \to \infty$, $\lambda^m \to \infty$.

Lemma 11. Let $f_m(\beta)$ denote the objective function of model F_λ as defined in (3.9), with arrival rate $\lambda^m = m\lambda^0$. And, let $g_m(\beta)$ denote the objective function of model G_λ as defined in (3.10), with arrival rate λ^m. Then,

$$\lim_{m\to\infty} \sup_{\beta\geq 0} (g_m(\beta) - f_m(\beta)) = 0.$$

Proof. To prove the desired result it suffices to prove

$$\prod_{i=1}^{L} (1 - UB(\beta_i, \lambda^m_i)) - \prod_{i=1}^{L} (1 - \tilde{\alpha}_\beta (\beta_i, \lambda^m_i))$$

converges uniformly to 0 in β as m goes to ∞. Let $\underline{UB}(\beta_i, \lambda^m_i)$ denote $1 - UB(\beta_i, \lambda^m_i)$ and let $\underline{\tilde{\alpha}_\beta}(\beta_i, \lambda^m_i)$ denote $1 - \tilde{\alpha}_\beta (\beta_i, \lambda^m_i)$. We will use mathematical induction to prove this lemma. We first prove

$$\underline{UB}(\beta_1, \lambda^m_1)\underline{UB}(\beta_2, \lambda^m_2) - \underline{\tilde{\alpha}_\beta}(\beta_1, \lambda^m_1)\underline{\tilde{\alpha}_\beta}(\beta_2, \lambda^m_2)$$

converges uniformly to 0 in $\beta = (\beta_1, \beta_2)$ as m goes to ∞. From Theorem 8, we have that this result holds separately for $UB(\beta_1, \lambda^m_1) - \tilde{\alpha}_\beta(\beta_1, \lambda^m_1)$ and

$UB(\beta_2, \lambda_2^m) - \tilde{\alpha}_\beta(\beta_2, \lambda_2^m)$, which implies that it again holds separately for $\underline{UB}(\beta_1, \lambda_1^m) - \underline{\tilde{\alpha}_\beta}(\beta_1, \lambda_1^m)$ and $\underline{UB}(\beta_2, \lambda_2^m) - \underline{\tilde{\alpha}_\beta}(\beta_2, \lambda_2^m)$. Also, since

$$\begin{aligned}
\underline{UB}(\beta_1, \lambda_1^m)\underline{UB}(\beta_2, \lambda_2^m) &- \underline{\tilde{\alpha}_\beta}(\beta_1, \lambda_1^m)\underline{\tilde{\alpha}_\beta}(\beta_2, \lambda_2^m) \\
&= \underline{UB}(\beta_1, \lambda_1^m)\underline{UB}(\beta_2, \lambda_2^m) - \underline{UB}(\beta_1, \lambda_1^m)\underline{\tilde{\alpha}_\beta}(\beta_2, \lambda_2^m) \\
&\quad + \underline{UB}(\beta_1, \lambda_1^m)\underline{\tilde{\alpha}_\beta}(\beta_2, \lambda_2^m) - \underline{\tilde{\alpha}_\beta}(\beta_1, \lambda_1^m)\underline{\tilde{\alpha}_\beta}(\beta_2, \lambda_2^m) \\
&= \underline{UB}(\beta_1, \lambda_1^m)\left(\underline{UB}(\beta_2, \lambda_2^m) - \underline{\tilde{\alpha}_\beta}(\beta_2, \lambda_2^m)\right) \\
&\quad + \left(\underline{UB}(\beta_1, \lambda_1^m) - \underline{\tilde{\alpha}_\beta}(\beta_1, \lambda_1^m)\right)\underline{\tilde{\alpha}_\beta}(\beta_2, \lambda_2^m).
\end{aligned}$$

Thus,

$$\begin{aligned}
&\lim_{m\to\infty}\sup_{\beta\geq 0}\left(\underline{UB}(\beta_1, \lambda_1^m)\underline{UB}(\beta_2, \lambda_2^m) - \underline{\tilde{\alpha}_\beta}(\beta_1, \lambda_1^m)\underline{\tilde{\alpha}_\beta}(\beta_2, \lambda_2^m)\right) \\
= &\lim_{m\to\infty}\sup_{\beta\geq 0}\left\{\underline{UB}(\beta_1, \lambda_1^m)\left(\underline{UB}(\beta_2, \lambda_2^m) - \underline{\tilde{\alpha}_\beta}(\beta_2, \lambda_2^m)\right)\right. \\
&\left. + \left(\underline{UB}(\beta_1, \lambda_1^m) - \underline{\tilde{\alpha}_\beta}(\beta_1, \lambda_1^m)\right)\underline{\tilde{\alpha}_\beta}(\beta_2, \lambda_2^m)\right\} \\
\leq &\lim_{m\to\infty}\sup_{\beta\geq 0}\left\{\underline{UB}(\beta_1, \lambda_1^m)\left(\underline{UB}(\beta_2, \lambda_2^m) - \underline{\tilde{\alpha}_\beta}(\beta_2, \lambda_2^m)\right)\right\} \\
&+ \lim_{m\to\infty}\sup_{\beta\geq 0}\left\{\left(\underline{UB}(\beta_1, \lambda_1^m) - \underline{\tilde{\alpha}_\beta}(\beta_1, \lambda_1^m)\right)\underline{\tilde{\alpha}_\beta}(\beta_2, \lambda_2^m)\right\}.
\end{aligned}$$

The above inequality holds because both terms are positive. Also we know both the Erlang-C formula and its upper bound have range [0, 1]. This indicates that

$$\lim_{m\to\infty}\sup_{\beta\geq 0}\left\{\underline{UB}(\beta_1, \lambda_1^m)\left(\underline{UB}(\beta_2, \lambda_2^m) - \underline{\tilde{\alpha}_\beta}(\beta_2, \lambda_2^m)\right)\right\} = 0$$

and

$$\lim_{m\to\infty}\sup_{\beta\geq 0}\left\{\left(\underline{UB}(\beta_1, \lambda_1^m) - \underline{\tilde{\alpha}_\beta}(\beta_1, \lambda_1^m)\right)\underline{\tilde{\alpha}_\beta}(\beta_2, \lambda_2^m)\right\} = 0.$$

Thus, we have

$$\lim_{m\to\infty}\sup_{\beta\geq 0}\left(\underline{UB}(\beta_1,\lambda_1^m)\underline{UB}(\beta_2,\lambda_2^m)-\underline{\tilde{\alpha}_\beta}(\beta_1,\lambda_1^m)\underline{\tilde{\alpha}_\beta}(\beta_2,\lambda_2^m)\right)=0.$$

Assume $\forall K\in\{2,3,\ldots,L-1\}$,

$$\prod_{i=1}^{K}\left(1-UB(\beta_i,\lambda_i^m)\right)-\prod_{i=1}^{K}\left(1-\tilde{\alpha}_\beta\left(\beta_i,\lambda_i^m\right)\right)$$

converges uniformly to 0 in β as m goes to ∞. We now prove that for $K+1$, we have

$$\prod_{i=1}^{K+1}\left(1-UB(\beta_i,\lambda_i^m)\right)-\prod_{i=1}^{K+1}\left(1-\tilde{\alpha}_\beta\left(\beta_i,\lambda_i^m\right)\right)$$

converges uniformly to 0 in β as m goes to ∞.

As above, we have

$$\begin{aligned}
&\left(\prod_{i=1}^{K}\left(1-UB(\beta_i,\lambda_i^m)\right)\right)\underline{UB}(\beta_{K+1},\lambda_{K+1}^m)-\left(\prod_{i=1}^{K}\left(1-\tilde{\alpha}_\beta\left(\beta_i,\lambda_i^m\right)\right)\right)\underline{\tilde{\alpha}_\beta}(\beta_{K+1},\lambda_{K+1}^m)\\
=\;&\left(\prod_{i=1}^{K}\left(1-UB(\beta_i,\lambda_i^m)\right)\right)\underline{UB}(\beta_{K+1},\lambda_{K+1}^m)-\left(\prod_{i=1}^{K}\left(1-UB(\beta_i,\lambda_i^m)\right)\right)\underline{\tilde{\alpha}_\beta}(\beta_{K+1},\lambda_{K+1}^m)\\
&+\left(\prod_{i=1}^{K}\left(1-UB(\beta_i,\lambda_i^m)\right)\right)\underline{\tilde{\alpha}_\beta}(\beta_{K+1},\lambda_{K+1}^m)-\left(\prod_{i=1}^{K}\left(1-\tilde{\alpha}_\beta\left(\beta_i,\lambda_i^m\right)\right)\right)\underline{\tilde{\alpha}_\beta}(\beta_{K+1},\lambda_{K+1}^m)\\
=\;&\left(\prod_{i=1}^{K}\left(1-UB(\beta_i,\lambda_i^m)\right)\right)\left(\underline{UB}(\beta_{K+1},\lambda_{K+1}^m)-\underline{\tilde{\alpha}_\beta}(\beta_{K+1},\lambda_{K+1}^m)\right)\\
&+\left(\prod_{i=1}^{K}\left(1-UB(\beta_i,\lambda_i^m)\right)-\prod_{i=1}^{K}\left(1-\tilde{\alpha}_\beta\left(\beta_i,\lambda_i^m\right)\right)\right)\underline{\tilde{\alpha}_\beta}(\beta_{K+1},\lambda_{K+1}^m),
\end{aligned}$$

Thus,

$$\begin{aligned}
&\lim_{m\to\infty}\sup_{\beta\geq 0}\left\{\prod_{i=1}^{K+1}\left(1-UB(\beta_i,\lambda_i^m)\right)-\prod_{i=1}^{K+1}\left(1-\tilde{\alpha}_\beta\left(\beta_i,\lambda_i^m\right)\right)\right\}\\
\leq\ &\lim_{m\to\infty}\sup_{\beta\geq 0}\left\{\left(\prod_{i=1}^{K}\left(1-UB(\beta_i,\lambda_i^m)\right)\right)\left(\underline{UB}(\beta_{K+1},\lambda_{K+1}^m)-\underline{\tilde{\alpha}_\beta}(\beta_{K+1},\lambda_{K+1}^m)\right)\right\}\\
&+\lim_{m\to\infty}\sup_{\beta\geq 0}\left\{\left(\prod_{i=1}^{K}\left(1-UB(\beta_i,\lambda_i^m)\right)-\prod_{i=1}^{K}\left(1-\tilde{\alpha}_\beta\left(\beta_i,\lambda_i^m\right)\right)\right)\underline{\tilde{\alpha}_\beta}(\beta_{K+1},\lambda_{K+1}^m)\right\}.
\end{aligned}$$

We know that $\left(\prod_{i=1}^{K}\left(1-UB(\beta_i,\lambda_i^m)\right)\right)$ and $\underline{\tilde{\alpha}_\beta}(\beta_{K+1},\lambda_{K+1}^m)$ are between 0 and 1. Then by Theorem 8 and the induction assumption, we have

$$\lim_{m\to\infty}\sup_{\beta\geq 0}\left\{\left(\prod_{i=1}^{K}\left(1-UB(\beta_i,\lambda_i^m)\right)\right)\left(\underline{UB}(\beta_{K+1},\lambda_{K+1}^m)-\underline{\tilde{\alpha}_\beta}(\beta_{K+1},\lambda_{K+1}^m)\right)\right\}=0$$

and

$$\lim_{m\to\infty}\sup_{\beta\geq 0}\left\{\left(\prod_{i=1}^{K}\left(1-UB(\beta_i,\lambda_i^m)\right)-\prod_{i=1}^{K}\left(1-\tilde{\alpha}_\beta\left(\beta_i,\lambda_i^m\right)\right)\right)\underline{\tilde{\alpha}_\beta}(\beta_{K+1},\lambda_{K+1}^m)\right\}=0.$$

This implies that

$$\prod_{i=1}^{K+1}\left(1-UB(\beta_i,\lambda_i^m)\right)-\prod_{i=1}^{K+1}\left(1-\tilde{\alpha}_\beta\left(\beta_i,\lambda_i^m\right)\right)$$

converges uniformly to 0 in β as m goes to ∞, establishing the result. □

Theorem 12. Let $f_m(\cdot)$ denote the objective function and let β_m^F denote the optimal solution of model F_λ as defined in (3.9), with arrival rate λ^m. Let $g_\lambda(\cdot)$ denote the objective function, and let β_m^G denote the optimal solution of model G_λ as defined in (3.10), with arrival rate λ^m. Then

$$\lim_{m\to\infty}\left(f_m(\beta_m^G)-f_m(\beta_m^F)\right)=0.$$

Proof. First, since β_m^F is optimal, with respect to F_λ with arrival rate λ^m, we have $f_m(\beta_m^G) - f_m(\beta_m^F) \geq 0$. The objective function g_λ is an upper bound for f_λ and so we have $g_m(\beta_m^G) \geq f_m(\beta_m^G)$. Thus $f_m(\beta_m^G) - f_m(\beta_m^F) \leq g_m(\beta_m^G) - f_m(\beta_m^F)$. We also have $g_m(\beta_m^F) \geq g_m(\beta_m^G)$, since β_m^G is optimal, with respect to G_λ with arrival rate λ^m. This implies that $g_m(\beta_m^G) - f_m(\beta_m^F) \leq g_m(\beta_m^F) - f_m(\beta_m^F)$. According to Lemma 11, we have $g_m(\beta_m^F) - f_m(\beta_m^F) \to 0$, as $m \to \infty$. This proves that $f_m(\beta_m^G) - f_m(\beta_m^F) \to 0$, as $m \to \infty$. □

Theorem 12 indicates that for the multi-station problem we describe above, our approximate solution asymptotically converges to the actual solution in the sense that for any $\delta > 0$, the gap between the optimal objective value and the objective value obtained by substituting our approximate solution into the objective function goes to 0 as $m \to \infty$.

We could also build the approximating problem using the lower bound on the Erlang-C formula, $LB(\beta, \lambda)$, to replace the Erlang-C formula in the original model F_λ. In this case, we can again obtain a result analogous to Theorem 12. We prefer to employ the upper bound rather than the lower bound because the solution under the former approximation is appropriately conservative, i.e., it is guaranteed to be feasible for model F_λ, while the solution under the lower bound is not.

Chapter 4

Stochastic Arrival-rate Problems

In this chapter, we extend the single-queue and the multi-queue models from Chapter 3 to model doubly stochastic service center systems. That is, we focus on solving large-scale service center staffing problems when the arrival rates are uncertain in addition to the inherent randomness of the system's inter-arrival times and service times. This brings the modeling closer to reality.

We provide a solution procedure for solving a staffing problem for a doubly stochastic service center system. We consider a decision making scheme in which we must select staffing levels before observing the arrival rates. We assume that the decision maker has distributional information about the arrival rates at the time of decision making. In the presence of arrival rate uncertainty, the decision maker's goal is to minimize the staffing cost, while ensuring the QoS achieves a given level.

We show that as the system scales large in size, there is at most one *key* scenario under which the probability of waiting converges to a non-trivial value, i.e., a value strictly between 0 and 1. In any other scenario, the probability of waiting converges to either 0 or 1, that is the staffing level is either over- or under-loaded in any scenario other than the key scenario, as the size

of the system grows to infinity. Exploiting this result, we propose a two-step solution procedure for the staffing problem with random arrival rates. In the first step, we use the desired QoS level to identify the key scenario corresponding to the optimal staffing level. After finding the key scenario, the random arrival-rate model reduces to a deterministic arrival-rate model. In the second step, we solve the resulting model, with deterministic arrival rate, by using our approximation model proposed in Chapter 3. The approximate optimal staffing level obtained in this procedure converges to the true optimal staffing level for the random arrival-rate problem as the system's size grows large.

4.1 Single-station System

We now extend our development to systems in which the arrival *rates* are random. We still consider a call center that has a single $M/M/n$ queue representing a single-station system in this section and then turn to a system with L parallel $M/M/n$ queues representing a multi-station system in the next section. The single-queue system and the multi-queue system that we consider in this chapter are depicted in Figures 4.1 and 4.2. The systems are similar to the ones we consider in Chapter 3, except that the arrival rates are random in this chapter.

First, we consider single-station system. Let Λ denote the random arrival rate. Let Λ^ω be a specific realization, where ω is a sample point from the sample space Ω. We assume that the sample space Ω is finite. Let p^ω be the probability of realizing scenario ω. Our first attempt to extend model

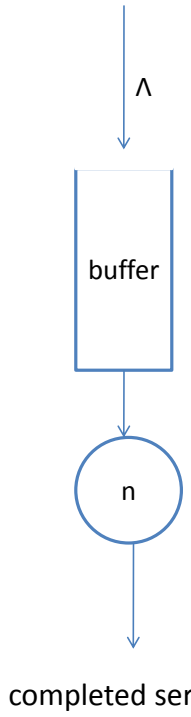

Figure 4.1: Single-Queue System - Stochastic Arrival Rate

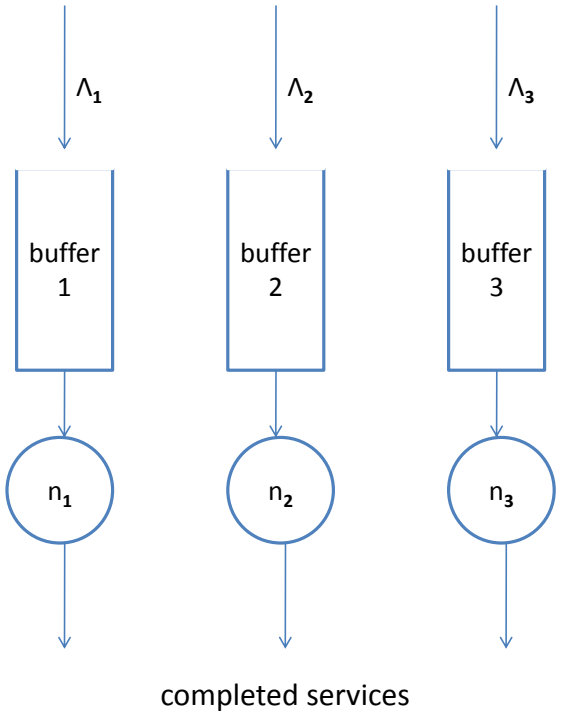

Figure 4.2: Multi-Queue System - Stochastic Arrival Rates

(3.3) to the doubly stochastic setting is as follows:

$$\min_{\beta \geq 0} \quad c(\beta) \qquad \text{s.t.} \sum_{\omega \in \Omega} p^{\omega} \mathbb{P}\left\{wait(\beta, \Lambda) > 0 \mid \Lambda = \Lambda^{\omega}\right\} \leq \epsilon, \tag{4.1}$$

where $\mathbb{P}\left\{wait(\beta, \Lambda) > 0 \mid \Lambda = \Lambda^{\omega}\right\} = \tilde{\alpha}_{\beta}(\beta, \Lambda^{\omega})$ and $\tilde{\alpha}_{\beta}(\beta, \Lambda^{\omega}) = \tilde{\alpha}(\Lambda^{\omega} + \beta\sqrt{\Lambda^{\omega}}, \Lambda^{\omega})$, where $\tilde{\alpha}(\cdot, \cdot)$ is defined in equation (2.1).

However, there is a fundamental shortcoming of this attempt. In a stochastic program a decision made "now" must be nonanticipative; i.e., it cannot depend on a realization of the randomness not yet observed. However, in this formulation, the number of servers does depend on ω in that the number of servers varies by ω via $\Lambda^{\omega} + \beta\sqrt{\Lambda^{\omega}}$. In order to rectify this, our decision must not only be β but also a specific scenario ω^{key} so that the number of servers is given by $\Lambda^{\omega^{key}} + \beta\sqrt{\Lambda^{\omega^{key}}}$ and does not vary by ω.

So, we revise the extension of model (3.3) as follows and denote the model F_{Λ}:

$$\min_{\beta \geq 0, \omega^{key} \in \Omega} c(\beta, \omega^{key}) \qquad \text{s.t.} \sum_{\omega \in \Omega} p^{\omega} \mathbb{P}\left\{wait(\beta, \Lambda) > 0 \mid \Lambda = \Lambda^{\omega}\right\} \leq \epsilon, \tag{4.2}$$

where $\mathbb{P}\{wait(\beta, \Lambda) > 0 | \Lambda = \Lambda^{\omega}\} \equiv \tilde{\alpha}(\Lambda^{\omega^{key}} + \beta\sqrt{\Lambda^{\omega^{key}}}, \Lambda^{\omega})$, where $\tilde{\alpha}(\cdot, \cdot)$ is defined in equation (2.1).

As in our previous development we are interested in an asymptotic result, as the arrival rate grows large. However, we now have realizations $\{\Lambda^{\omega}\}_{\omega \in \Omega}$ and we need to clarify what it means for Λ to grow large. We let Λ grow in the following way: We assume there is an initial value of the arrival rate in all scenarios. Let this value be $\Lambda_0 = (\Lambda_0^{\omega_1}, \ldots, \Lambda_0^{\omega_{|\Omega|}})$, where, without loss of

generality, we assume the components of Λ_0 satisfy $\Lambda_0^{\omega_1} < \Lambda_0^{\omega_2} < \cdots < \Lambda_0^{\omega_{|\Omega|}}$. Indexing the sequence of systems under consideration with the integers, we assume the arrival rate for the m^{th} system to be $\Lambda_m = m\Lambda_0$. Then, as $m \to \infty$, the arrival rate $\Lambda_m \to \infty$. We define (4.2) with Λ_m as F_{Λ_m}:

$$\min_{\beta \geq 0, \omega^{key} \in \Omega} \quad c(\beta, \omega^{key}) \qquad \text{s.t.} \sum_{\omega \in \Omega} p^{\omega} \mathbb{P}\left\{wait(\beta, \Lambda_m) > 0 \mid \Lambda_m = \Lambda_m^{\omega}\right\} \leq \epsilon. \quad (4.3)$$

The objective function $c(\beta, \omega)$ is assumed to have the following properties: $c(\cdot, \omega)$ is strictly increasing for all $\omega \in \Omega$, and $c(\beta, \omega_i) < c(\beta', \omega_{i+1})$ for all $\beta, \beta' \geq 0$.

As characterized in Theorem 1, when λ is deterministic, the probability of waiting has a non-degenerate limit if and only if the number of servers, n, increases in such a way that $n = \lambda + \beta\sqrt{\lambda}$ for some $\beta > 0$. In model (4.3), the number of servers n, or equivalently (β, ω^{key}), is chosen before we see the realization of the arrival rate. For a given staffing level, (β, ω^{key}), scenario ω^{key} is the only scenario for which the limit of the probability of waiting in this scenario is strictly between 0 and 1. In other scenarios, the system is either over- or under-loaded for the chosen staffing level as the arrival rate grows large. Thus we obtain a corollary of Theorem 1.

Corollary 13. For a given staffing level specified by $\beta > 0$ and ω^{key}, we have

$$\lim_{m \to \infty} \tilde{\alpha}(\Lambda_m^{\omega^{key}} + \beta\sqrt{\Lambda_m^{\omega^{key}}}, \Lambda_m^{\omega^{key}}) \in (0, 1).$$

For all $\omega \in \Omega$ such that $\omega \neq \omega^{key}$ and $\Lambda_0^{\omega} > \Lambda_0^{\omega^{key}}$, we have

$$\lim_{m \to \infty} \tilde{\alpha}(\Lambda_m^{\omega^{key}} + \beta\sqrt{\Lambda_m^{\omega^{key}}}, \Lambda_m^{\omega}) = 1;$$

for all $\omega \in \Omega$ such that $\omega \neq \omega^{key}$ and $\Lambda_0^{\omega} < \Lambda_0^{\omega^{key}}$, we have

$$\lim_{m\to\infty} \tilde{\alpha}(\Lambda_m^{\omega^{key}} + \beta\sqrt{\Lambda_m^{\omega^{key}}}, \Lambda_m^{\omega}) = 0.$$

Consider the constraint of model (4.3), for a specific decision $\beta > 0$ and $\omega_i = \omega^{key}$. Then, we have

$$\lim_{m\to\infty} \tilde{\alpha}(\Lambda_m^{\omega_i} + \beta\sqrt{\Lambda_m^{\omega_i}}, \Lambda_m^{\omega_i}) \in (0, 1). \tag{4.4}$$

Then, we approximate the QoS constraint in (4.3) via

$$\sum_{\omega=\omega_{i+1}}^{\omega_{|\Omega|}} p^{\omega} + p^{\omega_i}\mathbb{P}\left\{wait(\beta, \Lambda_m) > 0 \mid \Lambda_m = \Lambda_m^{\omega_i}\right\} \leq \epsilon. \tag{4.5}$$

This approximation replaces $\mathbb{P}\left\{wait(\beta, \Lambda_m) > 0 \mid \Lambda_m = \Lambda_m^{\omega}\right\}$ by unity for $\omega = \omega_{i+1}, \ldots, \omega_{|\Omega|}$, and by zero for $\omega = \omega_1, \ldots, \omega_{i-1}$. In view of Corollary 13 and equation (4.4), this approximation becomes increasingly precise as m grows large.

Equation (4.5) and the structure of $c(\beta, \omega_i)$ suggest that for sufficiently large m we should select the key scenario by finding the scenario ω_i such that $\sum_{k=i}^{|\Omega|} p^{\omega_k} \geq \epsilon$ and $\sum_{k=i+1}^{|\Omega|} p^{\omega_k} < \epsilon$, for $i \in \{1, 2, \ldots, |\Omega| - 1\}$; if $p^{\omega_{|\Omega|}} \geq \epsilon$, then we select $\omega_{|\Omega|}$ as the key scenario. In our work, we do not consider the trivial situations where $\epsilon = 0$ or $\epsilon = 1$. Thus, this mechanism for selecting the key scenario yields a unique ω_i. Given the key scenario ω_i, we form a first approximation to model (4.3) as:

$$\min_{\beta\geq 0} \quad c(\beta, \omega_i) \qquad \text{s.t. } p^{\omega_i}\mathbb{P}\left\{wait(\beta, \Lambda_m^{\omega_i}) > 0\right\} \leq \left(\epsilon - \sum_{k=i+1}^{|\Omega|} p^{\omega_k}\right). \tag{4.6}$$

The term $\mathbb{P}\{wait\,(\beta, \Lambda_m^{\omega_i}) > 0\}$ in model (4.6) is calculated by the Erlang-C formula, $\tilde{\alpha}(\Lambda_m^{\omega_i} + \beta\sqrt{\Lambda_m^{\omega_i}}, \Lambda_m^{\omega_i})$. We can use the upper bound function, $UB(\beta, \Lambda_m^{\omega_i})$, to approximate $\mathbb{P}\{wait(\beta, \Lambda_m^{\omega_i}) > 0\}$ and build our approximating model with ω_i which we denote G_{Λ_m}:

$$\min_{\beta \geq 0} \quad c(\beta, \omega_i) \qquad \text{s.t. } p^{\omega_i} UB\,(\beta, \Lambda_m^{\omega_i}) \leq \left(\epsilon - \sum_{k=i+1}^{|\Omega|} p^{\omega_k} \right). \tag{4.7}$$

Now we give a theorem by Buchanan and Hildebrandt [12] before we extend Theorem 10 in Chapter 3 and obtain Lemma 15.

Theorem 14. *(Buchanan and Hildebrandt [12])* If a sequence $f_n(x)$ of monotonic functions converges to a continuous function $f(x)$ in $[a, b]$ then this convergence is uniform.

Lemma 15. Let $\lambda > 0$. We extend model (3.5) in Chapter 3 to

$$\min_{\beta \geq 0} \quad c(\beta) \qquad \text{s.t. } \tilde{\alpha}_\beta(\beta, \lambda) \leq \epsilon_\lambda, \tag{4.8}$$

and denote its optimal solution by β_λ^F. We also extend model (3.6) in Chapter 3 to

$$\min_{\beta \geq 0} \quad c(\beta) \qquad \text{s.t. } UB(\beta, \lambda) \leq \epsilon_\lambda, \tag{4.9}$$

and denote its optimal solution by β_λ^G. Here the right-hand side ϵ_λ satisfies $\lim_{\lambda \to \infty} \epsilon_\lambda = \epsilon > 0$. Then $\beta_\lambda^G \geq \beta_\lambda^F$, $\forall \lambda > 0$, and there exists a finite β^* such that

$$\lim_{\lambda \to \infty} \beta_\lambda^G = \lim_{\lambda \to \infty} \beta_\lambda^F = \beta^*.$$

Proof. Denote the Halfin and Whitt approximation defined in (2.3) in Chapter 2 as $\alpha_{HW}(\cdot)$, and its inverse function as $\alpha_{HW}^{-1}(\cdot)$. Function $\alpha_{HW}(\cdot)$ is strictly decreasing and this implies that $\alpha_{HW}^{-1}(\cdot)$ is strictly decreasing. For any $\lambda > 0$, we denote the inverse of $UB(\cdot, \lambda)$ as $UB_\lambda^{-1}(\cdot)$. Function $UB(\cdot, \lambda)$ is strictly decreasing for any $\lambda > 0$, and this implies that $UB_\lambda^{-1}(\cdot)$ is strictly decreasing for any $\lambda > 0$. By Janssen et al. [25], we have

$$\lim_{\lambda \to \infty} UB(\beta, \lambda) = \alpha_{HW}(\beta), \forall \beta > 0.$$

Together with the monotonicity of $UB(\beta, \lambda)$ in λ for any $\beta > 0$, we have

$$\lim_{\lambda \to \infty} UB_\lambda^{-1}(x) = \alpha_{HW}^{-1}(x), \forall x > 0.$$

Since $\lim_{\lambda \to \infty} \epsilon_\lambda = \epsilon > 0$, there exist $l, u > 0$, such that $0 < l \leq \epsilon_\lambda \leq u$ when λ is large enough. Also, we know that $\alpha_{HW}^{-1}(\cdot)$ is continuous function. Then by Theorem 14, we have

$$\lim_{\lambda \to \infty} \sup_{x > 0} |UB_\lambda^{-1}(x) - \alpha_{HW}^{-1}(x)| = 0.$$

This gives

$$\lim_{\lambda \to \infty} UB_\lambda^{-1}(\epsilon_\lambda) = \alpha_{HW}^{-1}(\epsilon),$$

which implies that the limit of β_λ^G exists and we denote it as β^*. Then analogous to the proof of Theorem 10 in Chapter 3, we can show that the limit of β_λ^F exists, and $\lim_{\lambda \to \infty} \beta_\lambda^G = \lim_{\lambda \to \infty} \beta_\lambda^F = \beta^*$. □

In the following theorem, we can use Lemma 15 to infer that the gap between the optimal solution for model (4.3) and the optimal solution for model (4.7) goes to 0 as the system size increases.

Theorem 16. Let (ω_m^F, β_m^F) be an optimal solution to model F_{Λ_m} as defined in (4.3). Let β_m^G be an optimal solution to model G_{Λ_m} as defined in (4.7). Assume ϵ is such that there exists i with $\sum_{k=i}^{|\Omega|} p^{\omega_k} > \epsilon$ and $\sum_{k=i+1}^{|\Omega|} p^{\omega_k} < \epsilon$, and assume $c(\cdot, \omega)$ is strictly increasing for all $\omega \in \Omega$, and $c(\beta, \omega_i) < c(\beta', \omega_{i+1})$ for all $\beta, \beta' \geq 0$. Then, there exists $\bar{m}$ such that for all $m \geq \bar{m}$ we have $\omega_m^F = \omega_i$. And, there exists $\beta^* > 0$ such that

$$\lim_{m\to\infty} \beta_m^G = \lim_{m\to\infty} \beta_m^F = \beta^*.$$

Proof. From the hierarchical structure of $c(\beta, \omega)$ it is clear that ω_m^F is the element of $\{\omega_1, \omega_2, \ldots, \omega_{|\Omega|}\}$ with the smallest index for which there exists $\beta > 0$ such that the constraint of model (4.3) is feasible. In what follows we use $\omega < \omega'$ to mean $\Lambda_0^\omega < \Lambda_0^{\omega'}$. The number of scenarios, $|\Omega|$, is finite, then from Corollary 13 we have

$$\lim_{m\to\infty} \max_{\omega,\omega'\in\Omega,\omega\neq\omega'} \min\left\{\tilde{\alpha}(\Lambda_m^{\omega'} + \beta\sqrt{\Lambda_m^{\omega'}}, \omega), 1 - \tilde{\alpha}(\Lambda_m^{\omega'} + \beta\sqrt{\Lambda_m^{\omega'}}, \omega)\right\} = 0.$$

Thus, given $\delta > 0$, there exists $\bar{m}$, such that for all $m \geq \bar{m}$

$$\max_{\omega\in\Omega,\omega<\omega_m^F} \left\{\tilde{\alpha}(\Lambda_m^{\omega_m^F} + \beta\sqrt{\Lambda_m^{\omega_m^F}}, \omega)\right\} \leq \delta,$$

and

$$\max_{\omega\in\Omega,\omega>\omega_m^F} \left\{1 - \tilde{\alpha}(\Lambda_m^{\omega_m^F} + \beta\sqrt{\Lambda_m^{\omega_m^F}}, \omega)\right\} \leq \delta.$$

Thus, for all $m \geq \bar{m}$, the left-hand side of the constraint in (4.3) is bounded above by

$$\left(\sum_{\omega\in\Omega,\omega<\omega_m^F} p^\omega\right)\delta + \sum_{\omega\in\Omega,\omega>\omega_m^F} p^\omega + p^{\omega_m^F}\tilde{\alpha}(\Lambda_m^{\omega_m^F} + \beta\sqrt{\Lambda_m^{\omega_m^F}}, \omega_m^F) \tag{4.10}$$

and bounded below by

$$\left(\sum_{\omega\in\Omega,\omega>\omega_m^F} p^{\omega}\right)(1-\delta) + p^{\omega_m^F}\tilde{\alpha}(\Lambda_m^{\omega_m^F} + \beta\sqrt{\Lambda_m^{\omega_m^F}}, \omega_m^F). \tag{4.11}$$

Then for any $m \geq \bar{m}$, if $\omega_m^F < \omega_i$, then (4.11) indicates a contradiction of feasibility of F_{Λ_m}. If $\omega_m^F > \omega_i$, (4.10) indicates that model (4.3) is feasible for all $m \geq \bar{m}$. However, model (4.3) is also feasible in the ω_i case for all $m \geq \bar{m}$. Hence, from the hierarchical structure of $c(\beta, \omega)$, model (4.3) is suboptimal for $m \geq \bar{m}$ if $\omega_m^F > \omega_i$. This indicates that there exists $\bar{m}$, such that for all $m \geq \bar{m}$ we have $\omega_m^F = \omega_i$. For all $m \geq \bar{m}$, with ω_i, we can re-write (4.3) to a deterministic model like following:

$$\min_{\beta\geq 0} \quad c(\beta, \omega_i) \qquad \text{s.t. } p^{\omega_i}\mathbb{P}\left\{wait(\beta, \Lambda_m^{\omega_i}) > 0\right\} \leq \left(\epsilon - \sum_{k=i+1}^{|\Omega|} p^{\omega_k}\right) + \delta_m. \tag{4.12}$$

Here $\lim_{m\to\infty}\delta_m = 0$. Apply Lemma 15, we have

$$\lim_{m\to\infty}\beta_m^G = \lim_{m\to\infty}\beta_m^F = \beta^*.$$

□

4.2 Multi-station Systems

We now consider a multi-station system, again assuming that there are L queues, operated in a conditionally independent fashion. We define $\Lambda = (\Lambda_1, \ldots, \Lambda_L)$ as the random arrival rate vector. Let Λ^{ω} be a specific realization, where $\omega = (\omega_1, \ldots, \omega_L)$ is a sample point from the finite sample space $\Omega = \Omega_1 \times \cdots \times \Omega_L$. Let p^{ω} be the probability that scenario ω is realized.

We consider the following model:

$$\min_{\beta\geq 0,\omega^{key}\in\Omega} \quad \sum_{i=1}^{L} c_i(\beta_i,\omega_i^{key}) + \sum_{\omega\in\Omega} p^{\omega}\mathbb{P}\left\{\bigcup_{i=1}^{L}\{wait_i(\beta_i,\Lambda_i)>0\}\middle|\,\Lambda=\Lambda^{\omega}\right\}\leq\epsilon, \tag{4.13}$$

which is equivalent to the following model that we denote as F_Λ:

$$\begin{aligned}\min_{\beta\geq 0,\omega^{key}\in\Omega} \quad & \sum_{i=1}^{L} c_i(\beta_i,\omega_i^{key}) \\ \text{s.t.} \quad & \sum_{\omega\in\Omega} p^{\omega}\prod_{i=1}^{L}\mathbb{P}\left\{wait_i(\beta_i,\Lambda_i)=0|\,\Lambda_i=\Lambda_i^{\omega_i}\right\}\geq 1-\epsilon,\end{aligned} \tag{4.14}$$

where

$$\mathbb{P}\left\{wait_i(\beta_i,\Lambda_i)=0|\Lambda_i=\Lambda_i^{\omega_i}\right\} = 1-\tilde{\alpha}(\Lambda_i^{\omega_i^{key}}+\beta_i\sqrt{\Lambda_i^{\omega_i^{key}}},\Lambda_i^{\omega_i}).$$

Facing this random arrival-rate model, we may think that instead of solving the joint model (4.14), it may be easier to solve several individual models. That is, we can treat the L queues individually, and solve for the optimal staffing policy of each queue by solving a set of individual models. For example, instead of solving model (4.14), we may consider solving the following set of individual models:

$$\begin{aligned}\min_{\beta_i\geq 0,\omega_i^{key}\in\Omega_i} \quad & c_i(\beta_i,\omega_i^{key}) \\ \text{s.t.} \quad & \sum_{\omega_i\in\Omega_i} p_i^{\omega_i}\mathbb{P}\left\{wait_i(\beta_i,\Lambda_i)=0\mid\Lambda_i=\Lambda_i^{\omega_i}\right\} \\ & \geq \sqrt[L]{1-\epsilon}, \quad i=1,2,\ldots,L,\end{aligned} \tag{4.15}$$

where $p_i^{\omega_i}$ is the marginal probability of scenario $\omega_i\in\Omega_i$.

When decomposing the joint model (4.14) into the models in (4.15), it is difficult to decide on the right-hand side of the constraint in each individual

problem. In (4.15), we set the right-hand side of each to be $\sqrt[L]{1-\epsilon}$ with the notion that each service level should be the same across the individual problems, when there is no evidence showing that one is more important and deserves a higher service level than the others. Though the individual method is computationally easier, the solution it provides may be poor as shown in the following example.

Example 1. Let $L = 2$ and consider the resulting $M/M/n$ system as shown in Figure 4.3. Suppose each queue has random arrival rate Λ_i, $i = 1, 2$. Assume there are two scenarios for the arrival rate of queue 1, high and low; and, there are three scenarios for the arrival rate of queue 2, high, medium and low. The joint probability distribution is given in Table 4.1. The realizations of Λ for each queue under each scenario are given in Table 4.2. We assume a linear cost. The cost coefficients are $c_1 = 5$, $c_2 = 3$, in units of \$ per server, and the service level threshold value is $\epsilon = 0.05$.

Table 4.1: Joint Probability for Example 1

$p^{(\omega_1,\omega_2)}$	$\omega_2 = high$	$\omega_2 = medium$	$\omega_2 = low$
$\omega_1 = high$	0.03	0.21	0.1
$\omega_1 = low$	0.01	0.17	0.48

Table 4.2: Arrival Rates for Example 1

	high	medium	low
Λ_1 for queue 1	450	NA	350
Λ_2 for queue 2	300	200	100

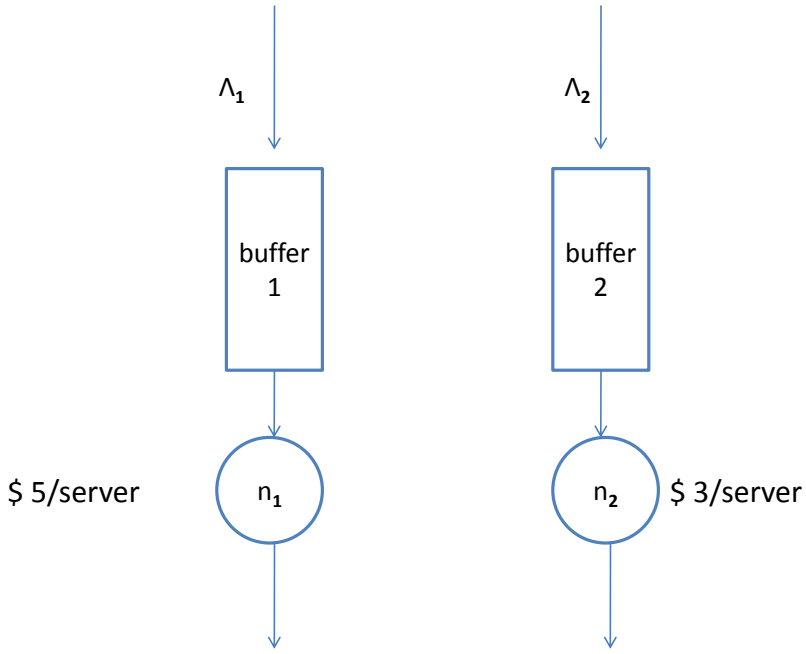

Figure 4.3: Depiction of the System for Example 1

We first provide the solutions to models (4.14) and (4.15) for the given parameters. We obtain the solutions that we detail in Table 4.3 by solving models (4.14) and (4.15), not the asymptotic versions of these models.

Table 4.3: Solution Comparison

	exact model (4.14)	individual models (4.15)
n	$(n_1^* = 496, n_2^* = 235)$	$(\overline{n}_1 = 484, \overline{n}_2 = 306)$
cost $(c_1 n_1 + c_2 n_2)$	3185	3338
$\mathbb{E}_\Lambda \left[\mathbb{P}\left\{\bigcup_{i=1}^{2} wait_i(n_i, \Lambda_i) > 0\right\}\right]$	0.05	0.05

From the solutions in Table 4.3 we can see that while both achieve the same service level, the staffing policy from the individual models (4.15) costs about 5% more than that of the joint model (4.14). □

We now discuss how to solve the joint model (4.14). Since our interest is in obtaining asymptotic results, we consider a sequence of queueing systems with increasing arrival rates. We let the arrival rates grow in the following way: We assume there is an initial value of the arrival rate in all scenarios. Let this initial rate be $\Lambda^0 = (\Lambda^{0\omega_1}, \ldots, \Lambda^{0\omega_{|\Omega|}})$, where each $\Lambda^{0\omega_k}, k = 1, \ldots, |\Omega|$, is an L-vector, since it represents the initial arrival rate the for L-queue system in scenario ω_k. Indexing the sequence of systems under consideration with the positive integers, we assume the arrival rate for the m^{th} system to be $\Lambda^m = m\Lambda^0$. Then, as $m \to \infty$, the arrival rate $\Lambda^m \to \infty$. Similar to the situation in our single-station system, we pick the staffing level for each queue $\beta_i, i = 1, \ldots, L$, before the realization of $\Lambda_i, i = 1, \ldots, L$. Thus as the arrival

rates grow, in at most one scenario, $\omega^{key} = (\omega_1^{key}, \ldots, \omega_L^{key})$, will the probability of waiting in each queue converge to a value strictly between 0 and 1. That is we can find a key scenario ω^{key}, such that $\forall \omega \neq \omega^{key}$, $\forall k \in \{1, \ldots, L\}$, if $\Lambda_k^{\omega_k} > \Lambda_k^{\omega_k^{key}}$, the limit probability of no waiting for queue k under this scenario is 0; and, if $\Lambda_k^{\omega_k} < \Lambda_k^{\omega_k^{key}}$, the limit probability of no waiting for queue k under this scenario is 1. Thus, once we find the key scenario, the random parameter model reduces to a deterministic model and the results in previous chapter can be applied. However, unlike the single-station system, we cannot easily identify the key scenario, for the same reason that a multivariate distribution does not have a unique quantile. To find the key scenario for multi-station system, we need to find the key scenario for each queue. We can identify the key scenario for each queue easily once we know how to allocate the ϵ to each queue. An integer programming model could be build to find the optimal allocation of the ϵ to each queue that minimizes the staffing costs. For any allocation of the ϵ, the existence of a key scenario for each queue is guaranteed by the definition of the key scenario for single-station system. This implies the existence of at least one key scenario for the multi-station system. Not like single-station system, multi-station system may have multiple key scenarios.

We use the data in Example 1 to explain our ideas in detail. It is obvious that the key scenario in this example is $(\omega_1, \omega_2) = (high, medium)$. Otherwise the largest possible value of the left-hand side of the constraint in model (4.14) cannot exceed $1 - \epsilon$. Also, with the key scenario being $(high, medium)$, we can select (β_1, β_2) to satisfy the constraint. Thus it is not necessary to check

scenario $(high, high)$, which is more costly. After finding the key scenario for Example 1, we can write the model (4.14) for the asymptotic version of the original random rates problem as:

$$\begin{aligned} \min_{\beta \geq 0} \quad & 5\beta_1 + 3\beta_2 \\ \text{s.t. } & 0.21\,(1 - \tilde{\alpha}_\beta(\beta_1, 450))\,(1 - \tilde{\alpha}_\beta(\beta_2, 200)) \\ & + 0.1(1 - \tilde{\alpha}_\beta(\beta_1, 450)) + 0.17(1 - \tilde{\alpha}_\beta(\beta_2, 200)) \\ & + 0.48 \geq (1 - 0.05). \end{aligned} \tag{4.16}$$

Solving model (4.16) we obtain the optimal solution $(\beta_1^*, \beta_2^*) = (2.15, 2.48)$, which gives

$$(n_1^*, n_2^*) = \Big(450 + 2.15 \cdot \sqrt{450}, \quad 200 + 2.48 \cdot \sqrt{200}\Big) \approx (496, 235).$$

We now solve the problem using the individual models (4.15). The individual models for queue 1 and queue 2 are:

$$\begin{aligned} \min_{\beta_1 \geq 0} \quad & 5\beta_1 \\ \text{s.t. } & 0.34(1 - \tilde{\alpha}_\beta(\beta_1, 450)) + 0.66(1 - \tilde{\alpha}_\beta(\beta_1, 350)) \geq \sqrt{1 - 0.05}, \end{aligned} \tag{4.17}$$

and

$$\begin{aligned} \min_{\beta_2 \geq 0} \quad & 3\beta_2 \\ \text{s.t. } & 0.04(1 - \tilde{\alpha}_\beta(\beta_2, 300)) + 0.38(1 - \tilde{\alpha}_\beta(\beta_2, 200)) \\ & + 0.58(1 - \tilde{\alpha}_\beta(\beta_2, 150)) \geq \sqrt{1 - 0.05}. \end{aligned} \tag{4.18}$$

The key scenarios for queue 1 and queue 2 are both $high$. The individual models above are equivalent to:

$$\min_{\beta_1 \geq 0} \quad 5\beta_1 \qquad \text{s.t. } 0.34(1 - \tilde{\alpha}_\beta(\beta_1, 450)) + 0.66 \geq \sqrt{1 - 0.05}, \tag{4.19}$$

and

$$\min_{\beta_2 \geq 0} \quad 3\beta_2 \qquad \text{s.t. } 0.04(1 - \tilde{\alpha}_\beta(\beta_2, 300)) + 0.38 + 0.58 \geq \sqrt{1 - 0.05}. \quad (4.20)$$

The optimal $\overline{\beta}_1$ and $\overline{\beta}_2$ achieved from solving (4.19) and (4.20) are $\overline{\beta}_1 = 1.6$, $\overline{\beta}_2 = 0.36$. This gives $\overline{n}_1 = 450 + 1.6 \cdot \sqrt{450} \approx 484$, $\overline{n}_2 = 300 + 0.36 \cdot \sqrt{300} \approx 306$.

In practice, model (4.14) can be hard to solve because of the complexity of the Erlang-C formula. As in the single-queue station system, after finding out the ω^{key}, we use the upper bound function $UB(\beta_i, \Lambda_i^\omega)$ to calculate $\mathbb{P}\{wait(\beta_i, \Lambda_i^\omega) > 0\}$ instead of using $\tilde{\alpha}_\beta(\beta_i, \Lambda_i^\omega)$, and we define the approximate model with ω^{key} as G_Λ:

$$\min_{\beta \geq 0, \omega = \omega^{key}} \quad \sum_{i=1}^{L} c_i(\beta_i, \omega^{key}) \qquad \text{s.t. } \sum_{\omega \in \Omega} p^\omega \left\{ \prod_{i=1}^{L} (1 - UB(\beta_i, \Lambda_i^\omega)) \right\} \geq 1 - \epsilon. \quad (4.21)$$

The following theorem characterizes asymptotic optimality of the multi-station system as the system size grows large.

Conjecture 17. Let $f_m(\cdot)$ denote the objective function and let (ω_m^F, β_m^F) denote the optimal solution of model F_Λ as defined in (4.14), with arrival rates $\Lambda^m = m\Lambda^0$. And, let β_m^G denote the optimal solution of model G_Λ as defined in (4.21), with arrival rates Λ^m. Then there exists $\bar{m}$ such that for all $m \geq \bar{m}$ we have $\omega_m^F = \omega^{key}$, and $\lim_{m \to \infty} \left(f_m(\beta_m^G) - f_m(\beta_m^F) \right) = 0$.

In this chapter, we describe a way to solve the random rate service center staffing problem. We first notice that as the arrival rate Λ grows, in at most one scenario does the probability that an arriving customer must wait

converge to a nontrivial value. This helps us to reduce the asymptotic version of the random rate problem to a deterministic rate problem after identifying the key scenario corresponding the optimal staffing level. After being reduced to a deterministic rate problem, we have, from the results in Chapter 3, that approximate solution solves the problem asymptotically.

Chapter 5

Two-stage Staffing Decision Problem

In the previous chapters, we focused on staffing decision over one decision time period. However, in the real world, the daily operation of a service center is split into hourly or half-hourly decision periods and the staffing decision needs to be made for each decision period. Also, when we consider the staffing problem with a random arrival rate, we assume that we know the distribution of the random arrival rate. However, the distribution of the arrival rate may vary over time and need to be updated based on new observations. In this chapter, we consider models that handle staffing decisions made over two adjacent decision periods (stages). We build models that minimize the staffing costs over two decision stages while satisfying a service quality constraint on the second stage operation. A Bayesian update is used to obtain the second-stage posterior arrival-rate distribution based on the first-stage prior arrival-rate distribution and the observations in the first stage. The second-stage distribution is used in the constraint on the second stage service quality. The problem considered in this chapter is a single-class single-station service center with random arrival rate. In the first section of this chapter, we assume the staffing decision for the first decision stage has been made, and focus on the relationship between the optimal second stage staffing decision and the obser-

vations from the first stage. In the second section of this chapter, we consider the situation where the first stage staffing decision is not given and needs to be made while taking into consideration the expected second stage staffing cost. A two-stage stochastic recourse formulation is built to analyze the relationship between the staffing decisions over the two periods. After reformulation, we show that our two-stage model can be rewritten as a newsvendor model. We then provide an algorithm which solves the two-stage staffing problem under several commonly used QoS constraints.

5.1 Two-stage Staffing Problem with Given First-stage Staffing Decision

We consider the problem of staffing a large-scale service center with a single class of customers and a single type of agent under a quality-of-service (QoS) constraint. The queueing model we use to represent such a service staffing problem is an $M/M/n$ model. We further assume the system we study has a stochastic *arrival rate*. That is, we assume that arrivals to the system occur according to a doubly stochastic Poisson process. In operating the service center over two time periods (stages), we assume that: (i) the distribution of the arrival rate for the first stage is known or has been previously estimated; (ii) the staffing level for the first stage, x_1, is given at the beginning of the first stage; and, (iii) the number of customers who arrive during the stage, n, is observed. We update the distribution of the arrival rate for the second stage based on n and then pick the staffing level, x_2, for stage two

based on the updated distribution. Figure 5.1 illustrates these time dynamics.

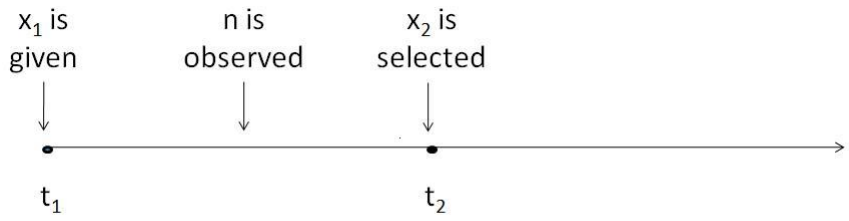

Figure 5.1: Time Dynamics of the Problem when x_1 is Given

5.1.1 Model Formulation

The service center's manager has two competing concerns. First the manager is concerned with the staffing cost for the second stage (we do not consider the cost for the first stage here, since the staffing level for the first stage is given), and hence would tend to hire as few servers in the second stage as possible. Second, the manager is concerned with service quality, which will be poor if an insufficient number of servers are hired. In this section, we use the function $\alpha(x_2, \lambda)$ to represent any quality service metric which depends on x_2 and λ, for example, this function could be the probability that a customer must wait, under a second period staffing level x_2 given arrival rate λ. We use Λ to denote the arrival rate as a random variable, and use λ to denote a deterministic value. Without loss of generality we assume that each server has unit service rate.

Let c be the unit staffing cost, c^+ be the unit staffing cost for additional

service agents, c^- be the unit salvage cost for sending unneeded service agents home and ϵ, which takes a value between the minimal and maximal possible values of service quality, be the service quality level threshold. Let $F_\Lambda(\lambda)$ be the CDF of the random arrival rate Λ, and $\alpha(x_2, \lambda)$ be the value of the QoS metric, conditioned on $\Lambda = \lambda$. The optimization model that minimizes staffing costs subject to the QoS constraint is then:

$$\min_{x_2 \geq 0} \quad cx_1 + c^+(x_2 - x_1)^+ - c^-(x_1 - x_2)^+ \tag{5.1a}$$

$$\text{s.t.} \quad \int_0^\infty \alpha(x_2, \lambda) dF_\Lambda(\lambda) \leq \epsilon. \tag{5.1b}$$

The integral in the QoS constraint in (5.1) simply gives the unconditional value of this QoS metric.

5.1.2 Gamma Prior Distribution

In our call volume forecasting model we assume that the prior distribution for Λ is $gamma(\alpha, \beta)$. The first period calls are then observed and used to produce an updated estimate for the distribution of Λ, i.e., the posterior distribution which is used in the second period. Since gamma is a conjugate prior when a Poisson likelihood function is used, the posterior distribution for Λ is also gamma. In particular, assume the prior distribution for the call volume Λ is $gamma(\alpha, \beta)$ with probability density function

$$g_1(\lambda_1; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda_1^{\alpha-1} e^{-\beta\lambda_1} \quad \text{for} \quad \lambda_1 \geq 0.$$

After observing n arrivals over $l \in \mathbb{R}_+$ minutes in the first stage, we obtain the estimated arrival rate distribution for the second stage (the posterior dis-

tribution), which is $gamma(\alpha+n, \beta+l)$ with density function

$$g_2(\lambda_2; n, \alpha, \beta, l) = \frac{(\beta+l)^{(\alpha+n)}}{\Gamma(\alpha+n)} \lambda_2^{\alpha+n-1} e^{-(\beta+l)\lambda_2} \quad \text{for} \quad \lambda_2 \geq 0.$$

To focus on the dependency of the second stage optimal staffing level on the number of observed arrivals in the first stage, in our problem, we assume l is fixed. Thus, to simplify the notation, we eliminate l from the parameter set of the posterior distribution, and denote its density function as $g_2(\lambda_2; n, \alpha, \beta)$ In this case, model (5.1) can be written as:

$$\min_{x_2 \geq 0} \quad cx_1 + c^+(x_2 - x_1)^+ - c^-(x_1 - x_2)^+ \tag{5.2a}$$

$$\text{s.t.} \quad \int_0^\infty g_2(\lambda; n, \alpha, \beta)\alpha(x_2, \lambda)d\lambda \leq \epsilon. \tag{5.2b}$$

Numerical Examples. To investigate the properties of the second-stage optimal solution, we solve the problem using various parameters in the prior distribution. Let $x_2^*(n; \alpha, \beta)$ denote the optimal second-stage staffing level as a function of n for the parameter set (α, β). In the experiments, we use the probability that a customer must wait for service as the service quality measurement. That is we assume

$$\alpha(x_2, \lambda_2) = \mathbb{P}(wait > 0 \mid x_2, \Lambda_2 = \lambda_2).$$

We use the Jagers-van Doorn continuous extension of the Erlang-C formula [24], that is

$$\alpha(x_2, \lambda_2) = \left[\lambda_2 \int_0^\infty te^{-\lambda_2 t}(1+t)^{x_2-1}dt\right]^{-1}.$$

The prior distribution for Λ is such that $\mathbb{E}\Lambda = \alpha/\beta$ and Var $\Lambda = \alpha/\beta^2$. In the experiments, when we vary α and β, we want them vary in such a way that α/β is fixed while α/β^2 is varied. The prior distribution is more concentrated about the mean $\mathbb{E}\Lambda = \alpha/\beta$ as the variance α/β^2 shrinks. Figure 5.2 shows the plot of $x_2^*(n;\alpha,\beta)$ versus n for different sets of (α,β). The figure depicts the solutions of (5.2) for parameter sets $(\alpha,\beta) = (2.5, 0.5)$, $(5,1)$, $(10,2)$, $(25,5)$. All the experiments in section 5.1 are performed on a PC with Intel Core Due CPU P9600 processors at 2.66GHz and 2.67GHz, and 2.00 GB of RAM. We summarize our observations on the numerical results shown in Figure 5.2 in the propositions and conjecture in the following paragraph.

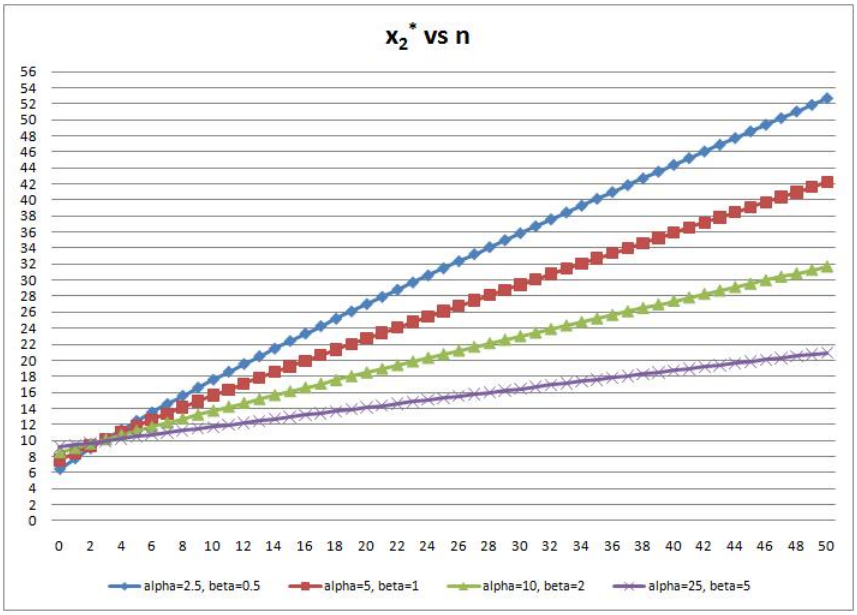

Figure 5.2: Function $x_2^*(n)$ for Gamma Prior Distribution

Characterizing Solutions. Define A as the subset of $\mathbb{R}_+^2$, on which the queueing system is stable. In most applications, an unstable system does not satisfy any reasonable QoS constraint. For example, suppose we consider the problem for a $M/M/n$ system and $\alpha(x, \lambda)$ is the probability a customer waits for service, then the system is only stable when $x > \lambda$. If $x < \lambda$, the stationary waiting time is infinite. Thus for the $M/M/n$ system, we consider quality measurement functions on set $A = \{(x, \lambda) \in \mathbb{R}_+^2 \bigcap \{x > \lambda > 0\}\}$. Before we state our results, we first give some conditions on the service quality measurement function $\alpha(x, \lambda) : A \to \mathbb{R}_+$,

(A1) $\alpha(x, \lambda)$ is a continuous function on A, and

$$\lim_{x \to \infty} \alpha(x, \lambda) = 0, \forall \lambda > 0,$$

and

$$\lim_{\lambda \to 0} \alpha(x, \lambda) = 0, \forall x > 0.$$

(A2) $\alpha(x, \lambda)$ is a continuous function on A, and $\alpha(x, \lambda)$ is strictly decreasing in x for any $\lambda > 0$ and strictly increasing in λ for any $x > 0$.

(A3) $\alpha(x, \lambda)$ is a continuous function on A, and $\alpha(x, \lambda)$ is differentiable in λ on A. $\frac{\partial \alpha(x,\lambda)}{\partial \lambda}$ is strictly decreasing in x for any $\lambda > 0$.

(A4) For any service quality level threshold ϵ,

$$\sup_{x>0} \alpha(x, \lambda) > \epsilon, \quad \forall \lambda > 0.$$

(A5) The distribution of Λ satisfies $\int_A dF_\Lambda(\lambda) > 0$.

Remark 1. Notice that $\alpha(x, \lambda)$ represents a QoS metric at arrival rate λ when we have x service agents. Our problem is a bi-criteria problem, the more service agents we have, the higher the staffing cost would be, and the lower the service quality would be. In our model, to control the service quality, we require $\alpha(x, \lambda)$ to be less than some pre-assigned threshold value ϵ in the constraint. Condition **(A1)** implies that when the arrival rate approaches 0, or when we have a large number of service agents, the service quality approaches the ideal level. Condition **(A2)** indicates that the service quality improves as the number of service agents increases, and deteriorates as the arrival rate increases. Thus, for most commonly used service quality measurements, such as the utilization, the continuous version of the probability a customer waits (given in [24]), and the continuous version of probability of abandonment (mentioned in (2.15) in Chapter 2), conditions **(A1)** and **(A2)** hold. Condition **(A4)** further guarantees the existence of the solution to model (5.1).

Remark 2. When condition **(A2)** holds, the function $\alpha(x, \lambda)$ is strictly increasing in λ for any $x > 0$. Condition (A3) is indicates that as more service agents are added, increased call volumes have a decreasing detrimental effect on the quality of service.

Proposition 18. Consider model (5.1) except replace the objective function with $C_{x_1}(x_2)$, where $C_{x_1}(x_2)$ is strictly increasing in x_2, and assume conditions **(A2)**, **(A4)** and **(A5)** hold for $\alpha(x_2, \lambda)$. Then there exists a unique solution to the associated model, denoted as x_2^*, where x_2^* solves $\int_0^\infty \alpha(x_2, \lambda) dF_\Lambda(\lambda) = \epsilon$.

Proof. Let $h(x_2) = \int_0^\infty \alpha(x_2, \lambda) dF_\Lambda(\lambda)$. We have $\alpha(x_2, \lambda)$ is strictly decreasing in x_2 for any $\lambda > 0$ on A. Thus **(A5)** implies that $h(x_2)$ is strictly decreasing in x_2. **(A4)** and the continuity of $\alpha(\cdot, \cdot)$ imply the existence of x_2^*. Since C_{x_1} is strictly increasing in x_2 and $\alpha(x_2, \lambda)$ is continuous, the solution to the optimization model is achieved at the boundary of the feasible region, that is, x_2^* is the solution to $\int_0^\infty \alpha(x_2, \lambda) dF_\Lambda(\lambda) = \epsilon$. Also x_2^* is unique, since $h(x_2)$ is strictly monotone in x_2. □

Remark 3. Note that by Proposition 18, x_2^* solves equation

$$\int_0^\infty \alpha(x_2, \lambda) dF_\Lambda(\lambda) = \epsilon$$

and hence does not depend on x_1.

Proposition 18 above can be applied to model (5.1) where function $\alpha(x, \lambda)$ may represent any QoS satisfying **(A2)**, **(A4)** and **(A5)**, and the arrival rate distribution need not be gamma. Proposition 19 and Conjecture 20 below are only for the specified model (5.2) in this section.

Proposition 19. Let $x_2^*(n; \alpha, \beta)$ denote the optimal solution to model (5.2) for the parameter set (α, β), given that n customers are observed in stage 1. Assume **(A1)** - **(A5)** hold for $\alpha(x_2, \lambda)$ and the shape parameter α, in the prior gamma distribution, is a positive integer. Then the optimal solution $x_2^*(n; \alpha, \beta)$ is a strictly increasing function of n for any fixed (α, β).

Proof. From Proposition 18, given fixed α, β, and n, $x_2^*(n; \alpha, \beta)$ solves

$$\int_0^\infty \alpha(x_2, \lambda) dF_\Lambda(\lambda; n, \alpha, \beta) = \epsilon,$$

and is unique. Also, we have

$$\begin{aligned}\int_0^\infty \alpha(x_2,\lambda)dF_\Lambda(\lambda;n,\alpha,\beta) &= \int_0^\infty \alpha(x_2,\lambda)d\mathbb{P}(\Lambda\le\lambda \mid n,\alpha,\beta)\\ &= \alpha(x_2,\lambda)\mathbb{P}(\Lambda\le\lambda \mid n,\alpha,\beta)|_0^\infty\\ &\quad - \int_0^\infty \mathbb{P}(\Lambda\le\lambda)d(\alpha(x_2,\lambda))\\ &= \int_0^\infty \frac{\partial\alpha(x_2,\lambda)}{\partial\lambda}d\lambda\\ &\quad - \int_0^\infty \frac{\partial\alpha(x_2,\lambda)}{\partial\lambda}\mathbb{P}(\Lambda\le\lambda \mid n,\alpha,\beta)d\lambda\\ &= \int_0^\infty \frac{\partial\alpha(x_2,\lambda)}{\partial\lambda}\mathbb{P}(\Lambda>\lambda \mid n,\alpha,\beta)d\lambda.\end{aligned}$$

The second-stage arrival rate follows gamma distribution with shape parameter $\alpha+n$ and scale parameter $\beta+l$. Since we assume α is an integer and n is the total number of arrivals in the first stage, which is also an integer, the shape parameter in the posterior distribution is still an integer. Let $G(\cdot \mid \alpha,\beta)$ be the CDF of gamma distribution. When the shape parameter can only take integer values, we have

$$G(\lambda \mid \alpha_1,\beta) > G(\lambda \mid \alpha_2,\beta), \quad \forall\lambda>0,\alpha_2>\alpha_1>0,\beta>0.$$

This implies that in the posterior distribution, $\mathbb{P}(\Lambda>\lambda \mid n,\alpha,\beta)$ is strictly increasing in n for any $\lambda>0$, that is

$$\mathbb{P}(\Lambda>\lambda \mid n_1,\alpha,\beta) < \mathbb{P}(\Lambda>\lambda \mid n_2,\alpha,\beta), \quad \forall n_1<n_2\in\mathbb{Z}_+.$$

Together with condition **(A3)**, we have that for $\forall n_1<n_2\in\mathbb{Z}_+$, $x_2^1<x_2^2$, where x_2^i satisfies

$$\int_0^\infty \frac{\partial\alpha(x_2^i,\lambda)}{\partial\lambda}\mathbb{P}(\Lambda>\lambda \mid n_i,\alpha,\beta)d\lambda=\epsilon, \quad i=1,2.$$

This implies our result. □

Remark 4. Notice that in Proposition 19, we require that the shape parameter of the prior distribution, α, take only integer values. We need this to get the dominance condition of the CDF of the posterior gamma distribution. In our application, the meaning of the shape parameter is the number of arrivals observed. Thus, it makes practical sense to assume that the initial shape parameter is a positive integer.

Conjecture 20. For any parameter sets (α_1, β_1) and (α_2, β_2), if $\frac{\alpha_1}{\beta_1} = \frac{\alpha_2}{\beta_2}$ and $\frac{\alpha_1}{{\beta_1}^2} < \frac{\alpha_2}{{\beta_2}^2}$, then $\frac{\partial x_2^*(n;\alpha_1,\beta_1)}{\partial n} < \frac{\partial x_2^*(n;\alpha_2,\beta_2)}{\partial n}$ for any $n > 0$.

Remark 5. If we fix the mean of the prior distribution while letting the variance of the prior distribution decrease, the prior distribution is then more concentrated around its mean. Conjecture 20 indicates that if this is the case, then the prior has more weight in the second stage staffing decision.

5.1.3 Discrete Prior Distribution

In the previous section, we assume the doubly stochastic Poisson process is governed by a gamma distribution. One may be tempted to simplify the problem by using a discrete distribution to model the arrival rate, so as to make the problem easier to solve. However, discretizing the distribution may result in badly behaved solutions, as demonstrated below. In this subsection, we assume that the arrival process is a doubly stochastic Poisson process with a discrete prior distribution for the first-stage arrival rate. For example, assume the first-stage arrival rate has a two-point discrete distribution

with probability mass function $\mathbb{P}\{\Lambda = \lambda_H\} = \mathbb{P}\{\Lambda = \lambda_L\} = 0.5$. If n arrivals are observed during the first stage, then we obtain the following arrival rate distribution for the second stage (the posterior distribution):

$$\mathbb{P}\{\Lambda = \lambda_H\} = \frac{e^{-\lambda_H}\lambda_H^n}{e^{-\lambda_H}\lambda_H^n + e^{-\lambda_L}\lambda_L^n}, \quad \mathbb{P}\{\Lambda = \lambda_L\} = \frac{e^{-\lambda_L}\lambda_L^n}{e^{-\lambda_H}\lambda_H^n + e^{-\lambda_L}\lambda_L^n}.$$

In this case, (5.1) can be written as:

$$\min_{x_2 \geq 0} \quad cx_1 + c^+(x_2 - x_1)^+ - c^-(x_1 - x_2)^+ \tag{5.3a}$$

$$\text{s.t.} \quad \frac{e^{-\lambda_H}\lambda_H^n}{e^{-\lambda_H}\lambda_H^n + e^{-\lambda_L}\lambda_L^n}\alpha(x_2, \lambda_H) + \frac{e^{-\lambda_L}\lambda_L^n}{e^{-\lambda_H}\lambda_H^n + e^{-\lambda_L}\lambda_L^n}\alpha(x_2, \lambda_L) \leq \epsilon. \tag{5.3b}$$

Numerical Example. We solve (5.3) for a two-point uniformly distributed arrival rate with $(\lambda_H, \lambda_L) = (70, 30)$ and the obvious generalization of (5.3) for a three-point uniformly distributed arrival rate with $(\lambda_H, \lambda_M, \lambda_L) = (70, 50, 30)$. In the experiments, we use the probability that a customer must wait for service as the service quality metric. Figure 5.3 shows the plot for the optimal second-stage staffing level, x_2^*, versus the number of arrivals observed in the first time stage, n. Even though using discrete distribution to approximate continuous distribution makes the problem easier to solve, but as can be seen in Figure 5.3, the optimal solutions from discrete models with a small sample space may be oversensitive to small changes in the observed call volume. Of course, if one sets up more bins in discretizing continuous distributions, the quality of results improve in the sense that the solution, x_2^*, will be less sensitive to changes in the observations. However, this advantage may be offset

by increased computational complexity. We summarize our observations on the numerical results shown in Figure 5.3 in the propositions in the following paragraph.

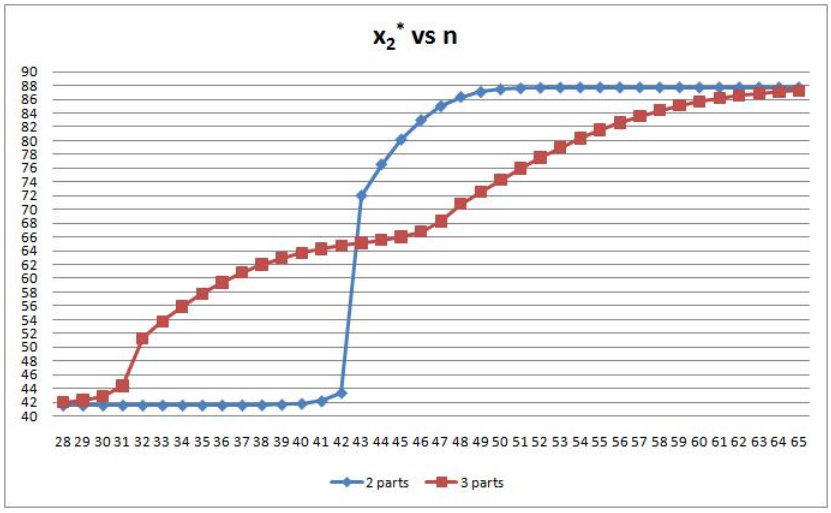

Figure 5.3: Function $x_2^*(n)$ for Discrete Prior Distribution

Characterizing Solutions.

Proposition 21. Assume conditions **(A1)**, **(A2)** and **(A4)** hold, then for a two-point discrete uniform prior distribution, the second-stage optimal solution $x_2^*(n)$ is strictly increasing in n.

Proof. By condition **(A1)** $\alpha(x_2, \lambda)$ is strictly decreasing in x_2 and the objective function of (5.3) is strictly increasing in x_2. Thus the optimal solution to (5.3) is the smallest $x_2 > 0$ such that

$$\frac{e^{-\lambda_H}\lambda_H^n}{e^{-\lambda_H}\lambda_H^n + e^{-\lambda_L}\lambda_L^n}\alpha(x_2, \lambda_H) + \frac{e^{-\lambda_L}\lambda_L^n}{e^{-\lambda_H}\lambda_H^n + e^{-\lambda_L}\lambda_L^n}\alpha(x_2, \lambda_L) \leq \epsilon.$$

As n increases,

$$\mathbb{P}\{\Lambda = \lambda_H \mid n\} \left(= \frac{e^{-\lambda_H}\lambda_H^n}{e^{-\lambda_H}\lambda_H^n + e^{-\lambda_L}\lambda_L^n}\right)$$

increases and

$$\mathbb{P}\{\Lambda = \lambda_L \mid n\} \left(= \frac{e^{-\lambda_L}\lambda_L^n}{e^{-\lambda_H}\lambda_H^n + e^{-\lambda_L}\lambda_L^n}\right)$$

decreases. Since $\mathbb{P}\{\Lambda = \lambda_H \mid n\} + \mathbb{P}\{\Lambda = \lambda_L \mid n\} = 1$ for any n, the amount of increase in $\mathbb{P}\{\Lambda = \lambda_H \mid n\}$ equals the amount of decrease in $\mathbb{P}\{\Lambda = \lambda_L \mid n\}$. However, for any x_2, we have

$$\alpha(x_2, \lambda_H) > \alpha(x_2, \lambda_L) \geq 0.$$

Thus $\forall n_1 < n_2 \in \mathbb{Z}_+$,

$$\begin{aligned}
&\frac{e^{-\lambda_H}\lambda_H^{n_2}}{e^{-\lambda_H}\lambda_H^{n_2} + e^{-\lambda_L}\lambda_L^{n_2}}\alpha(x_2, \lambda_H) - \frac{e^{-\lambda_H}\lambda_H^{n_1}}{e^{-\lambda_H}\lambda_H^{n_1} + e^{-\lambda_L}\lambda_L^{n_1}}\alpha(x_2, \lambda_H) > \\
&\frac{e^{-\lambda_L}\lambda_L^{n_1}}{e^{-\lambda_H}\lambda_H^{n_1} + e^{-\lambda_L}\lambda_L^{n_1}}\alpha(x_2, \lambda_L) - \frac{e^{-\lambda_L}\lambda_L^{n_2}}{e^{-\lambda_H}\lambda_H^{n_2} + e^{-\lambda_L}\lambda_L^{n_2}}\alpha(x_2, \lambda_L), \quad \forall x_2 > 0.
\end{aligned}$$

This proves that $\forall n_1 < n_2 \in \mathbb{Z}_+$, we have $x_2^*(n_1) < x_2^*(n_2)$, where $x_2^*(n_1)$ is the smallest $x_2 > 0$ such that

$$\frac{e^{-\lambda_H}\lambda_H^{n_1}}{e^{-\lambda_H}\lambda_H^{n_1} + e^{-\lambda_L}\lambda_L^{n_1}}\alpha(x_2, \lambda_H) + \frac{e^{-\lambda_L}\lambda_L^{n_1}}{e^{-\lambda_H}\lambda_H^{n_1} + e^{-\lambda_L}\lambda_L^{n_1}}\alpha(x_2, \lambda_L) \leq \epsilon,$$

and $x_2^*(n_2)$ is the smallest $x_2 > 0$ such that

$$\frac{e^{-\lambda_H}\lambda_H^{n_2}}{e^{-\lambda_H}\lambda_H^{n_2} + e^{-\lambda_L}\lambda_L^{n_2}}\alpha(x_2, \lambda_H) + \frac{e^{-\lambda_L}\lambda_L^{n_2}}{e^{-\lambda_H}\lambda_H^{n_2} + e^{-\lambda_L}\lambda_L^{n_2}}\alpha(x_2, \lambda_L) \leq \epsilon.$$

□

Proposition 22. Assume conditions **(A1)**, **(A2)** and **(A4)** hold. For a two-point discrete uniform prior distribution, if

$$\mathbb{P}\{\Lambda = \lambda_H \mid n = 0\} > \epsilon,$$

where ϵ is the RHS of the constraint in (5.3), then we have $x_2^*(n) > \lambda_H, \forall n \geq 0$; otherwise, there exists a "key" point in the number of arrivals during the first stage, denoted as n_{key}, such that $x_2^*(n) < \lambda_H$ for any $n \leq n_{key}$, and $x_2^*(n) > \lambda_H$ for any $n \geq n_{key} + 1$.

Proof. In Proposition 21, we prove that $x_2^*(n)$ is strictly increasing. If the posterior distribution has $\mathbb{P}\{\Lambda = \lambda_H \mid n\} > \epsilon$ even at $n = 0$, then x_2^* must be greater than λ_H to make the problem feasible, since otherwise, for the constraint in (5.3), we have

$$\begin{aligned}
&\frac{e^{-\lambda_H}\lambda_H^n}{e^{-\lambda_H}\lambda_H^n + e^{-\lambda_L}\lambda_L^n}\alpha(x_2, \lambda_H) + \frac{e^{-\lambda_L}\lambda_L^n}{e^{-\lambda_H}\lambda_H^n + e^{-\lambda_L}\lambda_L^n}\alpha(x_2, \lambda_L) \\
&\geq \frac{e^{-\lambda_H}\lambda_H^n}{e^{-\lambda_H}\lambda_H^n + e^{-\lambda_L}\lambda_L^n}\alpha(x_2, \lambda_H) \\
&= \frac{e^{-\lambda_H}\lambda_H^n}{e^{-\lambda_H}\lambda_H^n + e^{-\lambda_L}\lambda_L^n} = \mathbb{P}\{\Lambda = \lambda_H \mid n\} > \epsilon.
\end{aligned}$$

If the posterior distribution has $\mathbb{P}\{\Lambda = \lambda_H \mid n\} < \epsilon$ for all $n < \tilde{n}$ for some $\tilde{n} > 0$, then for small n ($n < \tilde{n}$), there exists an $\overline{x}_2(n) < \lambda_H$, such that

$$\frac{e^{-\lambda_H}\lambda_H^n}{e^{-\lambda_H}\lambda_H^n + e^{-\lambda_L}\lambda_L^n} + \frac{e^{-\lambda_L}\lambda_L^n}{e^{-\lambda_H}\lambda_H^n + e^{-\lambda_L}\lambda_L^n}\alpha(\overline{x}_2, \lambda_L) < \epsilon.$$

As n increases, in the posterior distribution, $\mathbb{P}\{\Lambda = \lambda_H \mid n\} = \frac{e^{-\lambda_H}\lambda_H^n}{e^{-\lambda_H}\lambda_H^n + e^{-\lambda_L}\lambda_L^n}$ increases. Thus when the observation n goes up to certain value, call it n_{key}

($n_{key} < \tilde{n}$), to satisfy the constraint, $\alpha(x_2^*, \lambda_H)$ needs to be less than one, that is the optimal solution x_2^* must be at least λ_H to make to problem feasible. □

Remark 6. Proposition 22 shows that if $\lambda_H - \lambda_L$ is large enough, the optimal second-stage solution x_2^* increases significantly when the number of arrivals observed in the first stage exceeds a certain value. Such a solution may be viewed as badly behaved in that the optimal number of servers is very sensitive to small changes in the observed data.

Proposition 23. Assume conditions **(A1)**, **(A2)** and **(A4)** hold. For a three-point discrete prior distribution, $x_2^*(n)$ is strictly increasing. If the

$$\mathbb{P}\{\Lambda = \lambda_H \mid n = 0\} > \epsilon,$$

then we have $x_2^*(n) > \lambda_H$, $\forall n \geq 0$. If the

$$\mathbb{P}\{\Lambda = \lambda_H \mid n = 0\} + \mathbb{P}\{\Lambda = \lambda_M \mid n = 0\} > \epsilon,$$

then we have $x_2^*(n) > \lambda_M$, $\forall n \geq 0$; otherwise, and there exist two "key" points in the number of arrivals during the first stage, denoted as n_{key_1} and n_{key_2}, such that $x_2^*(n) < \lambda_M$ for any $n \leq n_{key_1}$, and $x_2^*(n) > \lambda_M$ for any $n \geq n_{key_1} + 1$; and $x_2^*(n) < \lambda_H$ for any $n \leq n_{key_2}$, and $x_2^*(n) > \lambda_H$ for any $n \geq n_{key_2} + 1$.

Proof. Similar to the proofs of Propositions 21 and 22. □

5.2 Two-stage Staffing Problem

Now we start to consider the true two-stage problem. We extend the problem considered in the above section to a two-stage problem, in which

the first stage staffing decision, x_1, is also a decision variable. Similar to the problem previous section, we still consider the problem of staffing a service center with a single class of customers and a single type of agents under a quality-of-service (QoS) constraint. Again we assume that arrivals to the system occur according to a doubly stochastic Poisson process and the queueing model we use to represent the staffing problem is an $M/M/n$ model. Considering operating the service center over two time periods, we assume that: (i) the distribution of the arrival rate for the first stage is known or has been previously estimated; (ii) the staffing level for the first-stage, x_1, is selected at the beginning of the first stage; and, (iii) the number of customers who arrive during the first stage, n, is observed. We update the distribution of the arrival rate for the second stage based on n and then pick the staffing level, x_2, for stage two based on the updated distribution. Figure 5.4 illustrates these time dynamics.

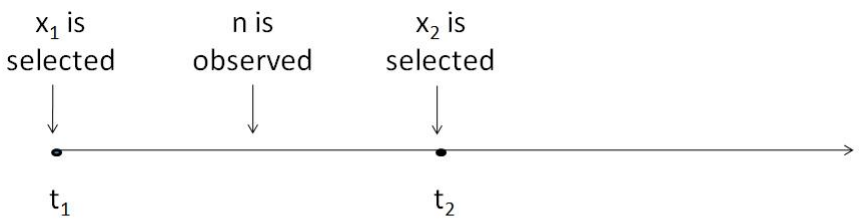

Figure 5.4: Time Dynamics of the Problem when x_1 is Optimized

5.2.1 Model Formulation

We start with the following general two-stage model. Let N denote the number of arrivals in the first stage, and let n represent a realization of N. Then the two-stage model is as follows:

$$\min_{x_1 \geq 0} \quad cx_1 + \mathbb{E}_N h(x_1, N), \tag{5.4a}$$

$$\text{where} \quad h(x_1, N|_{N=n}) = \min_{x_2 \geq 0} \quad c^+(x_2 - x_1)^+ - c^-(x_1 - x_2)^+ \tag{5.4b}$$

$$\text{s.t.} \quad \textit{QoS constraint.} \tag{5.4c}$$

It is obvious that the optimal second-stage staffing level x_2^* does not depend on the first-stage staffing level x_1, as long as the QoS constraint is only on the second-stage service quality. The second-stage optimal staffing level x_2^* is affected by the observation from the first-stage, N, since the posterior distribution of the arrival rate depends on N. So x_2^* is a function of N, and thus x_2^* is a random variable, which we denote by $x_2^*(N)$. The specific of function $x_2^*(N)$ is determined by the QoS constraint. The optimal second-stage cost, on the other hand, depends on the value of x_1. This means the optimal first-stage staffing level x_1^* is determined by the distribution of the optimal second-stage staffing level $x_2^*(N)$.

5.2.2 Two-stage Model with Constraint on Utilization

Now we describe in detail our two-stage model using utilization as the metric in the QoS constraint. In this case, the model is

$$\min_{x_1 \geq 0} \quad cx_1 + \mathbb{E}_N h(x_1, N), \tag{5.5a}$$

$$\text{where} \quad h(x_1, N|_{N=n}) = \min_{x_2 \geq 0} \quad c^+(x_2 - x_1)^+ - c^-(x_1 - x_2)^+ \tag{5.5b}$$

$$\text{s.t.} \quad \mathbb{P}_{\Lambda_2|_{N=n}} \left(\frac{\Lambda_2}{x_2} < \delta \right) \geq 1 - \epsilon. \tag{5.5c}$$

Here, ϵ and δ are some pre-selected values between 0 and 1.

It is obvious that in (5.5), $x_2^*(N)|_{N=n}$ is determined only by the second-stage constraint and we have

$$x_2^*(N)|_{N=n} \in \arg\min \left\{ x \geq 0 : \mathbb{P}_{\Lambda_2|_{N=n}} \left(\frac{\Lambda_2}{x} < \delta \right) \geq 1 - \epsilon \right\}.$$

As before, we assume that $\Lambda_1 \sim gamma(\alpha, \beta)$, and we use a Bayesian update to obtain the distribution of Λ_2 after observing N. That is, after observing n arrivals over $l \in \mathbb{R}_+$ minutes in the first stage, we have $\Lambda_2 \sim gamma(\alpha + n, \beta + l)$. Thus, using $F_{\Lambda_2|_{N=n}}(\cdot)$ to represent the CDF of the gamma distribution for Λ_2 given $N = n$, (5.2.2) becomes

$$x_2^*(N)|_{N=n} \in \arg\min \left\{ x : F_{\Lambda_2|_{N=n}}(\delta x) = \frac{\gamma(\alpha + n, (\beta + l)\delta x)}{\Gamma(\alpha + n)} \geq 1 - \epsilon, x \geq 0 \right\}.$$

Let $G_n(\cdot)$ be the CDF of a gamma distribution with parameters $\alpha + n$ and $(\beta + l)\delta$, then we have

$$F_{\Lambda_2|_{N=n}}(\delta x) = G_n(x)$$

and

$$x_2^*(N)|_{N=n} = \lceil G_n^{-1}(1-\epsilon) \rceil. \tag{5.6}$$

We re-write (5.5) with the optimal second stage staffing level, x_2^*:

$$\min_{x_1 \geq 0} \quad cx_1 + \mathbb{E}_N[c^+(x_2^*(N) - x_1)^+ - c^-(x_1 - x_2^*(N))^+]. \tag{5.7}$$

Model (5.7) can be re-written as:

$$\max_{x_1 \geq 0} \quad \mathbb{E}_N[-cx_2^*(N) - (c^+ - c)(x_2^*(N) - x_1)^+ - (c - c^-)(x_1 - x_2^*(N))^+]. \tag{5.8}$$

Model (5.8) has the form of a standard newsvendor's problem. Therefore, the solution is given by:

$$x_1^* \in \arg\min \left\{ x \geq 0 : \mathbb{P}_N(x_2^*(N) \leq x) \geq \frac{c^+ - c}{c^+ - c^-} \right\}.$$

Now, we discuss the distribution of $x_2^*(N)$. As mentioned before, x_2^* is a function of N. Thus to obtain the distribution of x_2^*, we need to obtain the distribution of N. Under our assumptions, $\Lambda_1 \sim gamma(\alpha, \beta)$, and $N \sim Poisson(\Lambda_1)$. Use $g(\lambda; \alpha, \beta)$ to stand for the PDF of a gamma distribution with parameters α and β, we have

$$\begin{aligned}
\mathbb{P}(N = n) &= \int_0^\infty g(\lambda; \alpha, \beta) \mathbb{P}(N = n | \Lambda_1 = \lambda) d\lambda \\
&= \int_0^\infty g(\lambda; \alpha, \beta) \frac{\lambda^n e^{-\lambda}}{n!} d\lambda \\
&= \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \frac{\lambda^n e^{-\lambda}}{n!} d\lambda \\
&= \frac{\beta^\alpha \Gamma(n+\alpha)}{\Gamma(\alpha) n! (\beta+1)^{\alpha+n}} \int_0^\infty \frac{\lambda^{\alpha+n-1} e^{-(\beta+1)\lambda} (\beta+1)^{\alpha+n}}{\Gamma(n+\alpha)} d\lambda \\
&= \frac{\beta^\alpha \Gamma(n+\alpha)}{\Gamma(\alpha) n! (\beta+1)^{\alpha+n}}.
\end{aligned}$$

Notice that in the above formula, if α is a positive integer, then

$$\mathbb{P}(N = n) = \binom{n+\alpha-1}{n} \left(\frac{1}{\beta+1}\right)^n \left(\frac{\beta}{\beta+1}\right)^\alpha.$$

This implies that N has a negative binomial distribution with parameters α and $\frac{\beta}{\beta+1}$, when α is an integer. That is $N \sim NegBin(\alpha, \frac{\beta}{\beta+1})$. There are a couple variations of the negative binomial distribution. Here, we are using the version of the negative binomial distribution that counts the number of failures before the αth success. With this version, the PMF of the negative binomial distribution is

$$\mathbb{P}(K = k|p, \alpha) = \binom{\alpha+k-1}{k} p^\alpha (1-p)^k, \quad k \in \mathbb{Z}_+.$$

It is possible to extend the definition of the negative binomial distribution to the case of a positive real parameter α. The PMF for this extended negative binomial distribution is

$$\mathbb{P}(K = k|p, \alpha) = \frac{\Gamma(k+\alpha)}{\Gamma(\alpha)k!} p^\alpha (1-p)^k, \quad k \in \mathbb{Z}_+.$$

Denote the CDF of the extended negative binomial distribution as

$$H(k; \alpha, p) = \mathbb{P}(K \leq k|p, \alpha) = \sum_{i=0}^{k} \frac{\Gamma(i+\alpha)}{\Gamma(\alpha)i!} p^\alpha (1-p)^i, \quad k \in \mathbb{Z}_+.$$

We have that $N \sim H(\cdot; \alpha, \frac{\beta}{\beta+1})$. We now provide the algorithm for solving the two-stage model (5.5) by summarizing the content in this section.

Now, we discuss about the above algorithm in detail. Let $F_{x_2^*(N)}(\cdot)$ be the CDF of $x_2^*(N)$, which is a càdlàg function. We define the generalized

step 1 **(find n^* corresponding to x_1^*)**
select $n^* \in \arg\min \left\{ k \in \mathbb{Z}_+ : H(k, \alpha, \frac{\beta}{\beta+1}) \geq \frac{c^+ - c}{c^+ - c^-} \right\}$;

step 2 **(obtain $x_1^* = x_2^*(N)|_{N=n^*}$)**
select $x_2^*(N)|_{N=n^*} \in$
$\arg\min \left\{ x : F_{\Lambda_2|_{N=n^*}}(\delta x) = \frac{\gamma(\alpha+n^*, (\beta+l)\delta x)}{\Gamma(\alpha+n^*)} \geq 1-\epsilon, x \geq 0 \right\}$;

inverse of $F_{x_2^*(N)}(\cdot)$. Let $F^{-1}_{x_2^*(N)}(y) = \inf_{x \in \mathbb{R}} \left\{ F_{x_2^*(N)}(x) \geq y \right\}$. We want to obtain the smallest x_1^* that satisfies

$$x_1^* \geq F^{-1}_{x_2^*(N)} \left(\frac{c^+ - c}{c^+ - c^-} \right),$$

or equivalently

$$F_{x_2^*(N)}(x_1^*) \geq \frac{c^+ - c}{c^+ - c^-}.$$

That is

$$\mathbb{P}(x_2^*(N) \leq x_1^*) \geq \frac{c^+ - c}{c^+ - c^-},$$

or equivalently

$$\mathbb{P}(N \leq x_2^{*-1}(x_1^*)) \geq \frac{c^+ - c}{c^+ - c^-}.$$

Denote $x_2^{*-1}(x_1^*)$ as n^*. In step 1, we solve for this n^*, and in step 2, we find x_1^* by evaluating function $x_2^*(n^*)$.

The experiments described in this paragraph show the results of solving (5.5) with various value of α and β using the algorithm described above. In the experiments, we fix α to be 900, and let β decrease from 45 to 10 with a unit decrement. In such a way, the coefficient of variation of the first-stage arrival rate, $\frac{\sqrt{var(\Lambda_1)}}{mean(\Lambda_1)}$, is fixed, while the mean of the first-stage arrival rate,

$mean(\Lambda_1)$, varies from 20 to 90. The cost parameters are set to be $c^+ = 4$, $c = 2$ and $c^- = 1$. The service quality threshold value, ϵ, is set to be 0.05. We conducted the experiments in MATLAB 7.11 (64 bit), and it took 0.22 seconds for MATLAB to finish the experiments. All the experiments in section 5.2 are performed on a PC with Intel Core i7-980 processors at 3.88GHz, and 24.00 GB of RAM.

We also conduct other two sets of experiments, in which we let α and β vary in the way that $mean(\Lambda_1)$ is fixed at 45, and the coefficient of variation of Λ_1 varies from 0.03 to 0.27. Figure 5.5 plots the optimal number of servers against the coefficient of variation of the first-stage arrival rate at $\frac{c^+-c}{c^+-c^-} = \frac{2}{3}$. Figure 5.6 plots the optimal number of servers against the coefficient of variation of the first-stage arrival rate at $\frac{c^+-c}{c^+-c^-} = \frac{2}{5}$. In both figures, we can see that as the COV of Λ_1 starts to increase, the optimal solutions also increase to cover the additional risk (variation). However, compare the two figures, in figure 5.5, the optimal number of servers increase faster than that in figure 5.6. This is because that relatively, in the model with $\frac{c^+-c}{c^+-c^-} = \frac{2}{3}$, the penalty for over-staffing is lower than the penalty for over-staffing in the model with $\frac{c^+-c}{c^+-c^-} = \frac{2}{5}$. Thus the solution for the model with $\frac{c^+-c}{c^+-c^-} = \frac{2}{3}$ tends to grow faster facing the increase in the COV of the randomness to cover the uncertainty in the random parameter.

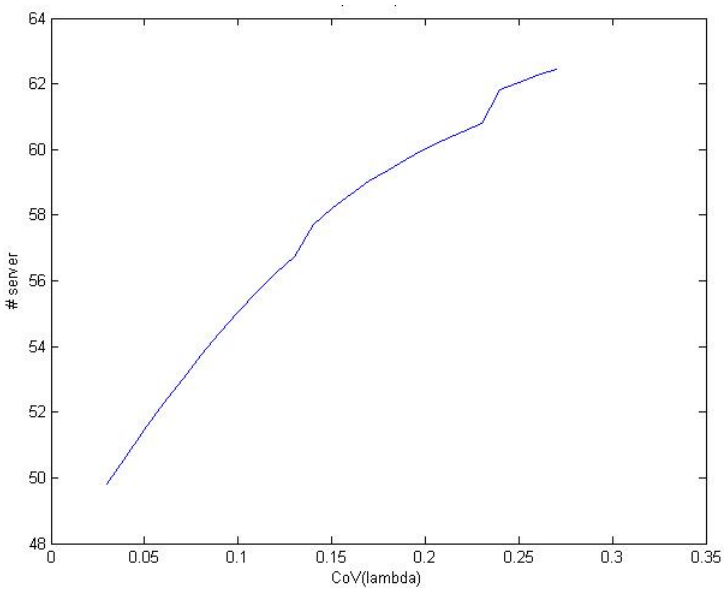

Figure 5.5: $x_1^*(n)$ vs $COV(\Lambda_1)$ for Utilization Model at $\frac{c^+-c}{c^+-c^-} = \frac{2}{3}$

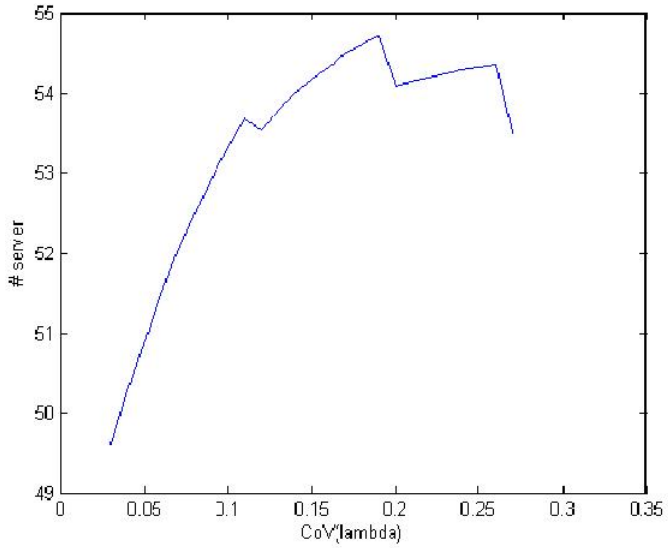

Figure 5.6: $x_1^*(n)$ vs $COV(\Lambda_1)$ for Utilization Model at $\frac{c^+-c}{c^+-c^-} = \frac{2}{5}$

5.2.3 Two-stage Model with Constraint on Probability of Waiting

As mentioned before, the QoS constraint can be of any type. When we use constraints other than the utilization constraint appearing in (5.5), step 1 is the same. However, in step 2, the function $x_2^*(\cdot)$, which is determined by the second-stage constraint in the model, is different, and the level of difficulty in solving the problem with other kinds of QoS constraints depends on the level of difficulty in evaluating the function $x_2^*(\cdot)$.

To illustrate the complexity introduced by applying other types of QoS constraints, we apply model (5.4) to an $M/M/n$ queueing system with a QoS constraint on the probability of waiting. In particular, we have

$$\min_{x_1 \geq 0} \quad cx_1 + \mathbb{E}_N h(x_1, N), \tag{5.9a}$$

$$\text{where} \quad h(x_1, N|_{N=n}) = \min_{x_2 \geq 0} \quad c^+(x_2 - x_1)^+ - c^-(x_1 - x_2)^+ \tag{5.9b}$$

$$\text{s.t.} \quad \mathbb{P}_{\Lambda_2|_{N=n}} \left(\mathbb{P}(wait > 0 | x_2, \Lambda_2) < \delta\right) \geq 1 - \epsilon. \tag{5.9c}$$

In solving (5.9), the only difference from solving (5.5) is the function of x_2^* of N. In (5.9), function $x_2^*(N)$ is determined by finding

$$x_2^*(N)|_{N=n} \in \arg\min \left\{x : \mathbb{P}_{\Lambda_2|_{N=n}} \left(\mathbb{P}(wait > 0 \mid x_2, \Lambda_2) < \delta\right) \geq 1 - \epsilon, x \geq 0\right\}.$$

Using the Jagers-van Doorn continuous extension of the Erlang-C formula [24] for $\mathbb{P}(wait > 0 \mid x_2, \Lambda_2)$, we have

$$x_2^*(N)|_{N=n} \in \arg\min \{x \geq 0 : \mathbb{P}_{\Lambda_2|_{N=n}} \left(\left[\Lambda_2 \int_0^\infty te^{-\Lambda_2 t}(1+t)^{x-1} dt\right]^{-1} < \delta \right) \geq 1 - \epsilon \Big\}.$$

Because of the complexity of the formula for $\mathbb{P}(wait > 0 \mid x_2, \Lambda_2)$, once we obtain n^* from step 1, it is not as easy as it is for the utilization constraint model to evaluate $x_2^*(n^*)$. Instead of the relativity explicit formula appearing in (5.6), one needs to apply a line search to perform this evaluation.

In the next set of experiments, we solved (5.9) with the same set of values on α and β as in the experiments for solving (5.5). The ϵ and δ in (5.9) are both set to be 0.05. As we mentioned above, line searches on x_2 are needed in step 2 of the algorithm. The lower and upper bounds of the line search are set to be 1 and 120, and the tolerance level of the line search is set to be 0.01. We conducted the experiments again in MATLAB 7.11 (64 bit), and it took 1054.91 seconds for MATLAB to finish the experiments.

Like in the utilization constraint model, we also conduct other two sets of experiments, in which we let α and β vary in the way that $mean(\Lambda_1)$ is fixed at 45, and the coefficient of variation of Λ_1 varies from 0.03 to 0.27. Figure 5.7 plots the optimal number of servers against the coefficient of variation of the first-stage arrival rate at $\frac{c^+-c}{c^+-c^-} = \frac{2}{3}$. Figure 5.8 plots the optimal number of servers against the coefficient of variation of the first-stage arrival rate at $\frac{c^+-c}{c^+-c^-} = \frac{2}{5}$. Like before, in both figures, we can see that as the COV of Λ_1 starts to increase, the optimal solutions also increase to cover the additional risk (variation). However, compare the two figures, in figure 5.7, the optimal number of servers increase faster than that in figure 5.8.

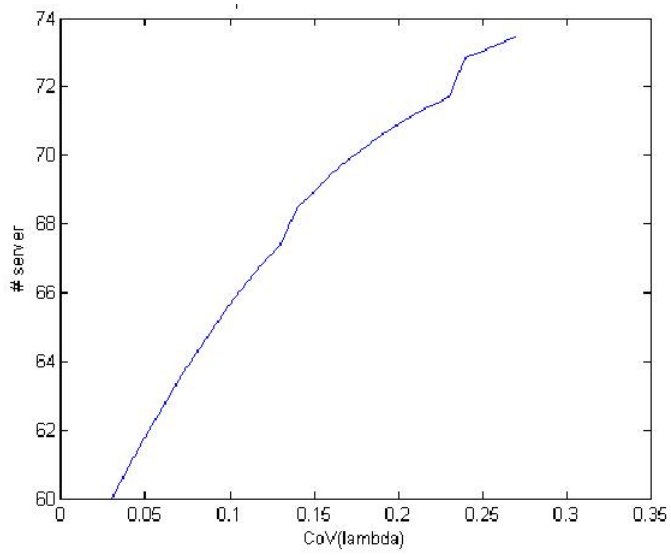

Figure 5.7: $x_1^*(n)$ vs $COV(\Lambda_1)$ for Probability of Waiting Model at $\frac{c^+-c}{c^+-c^-} = \frac{2}{3}$

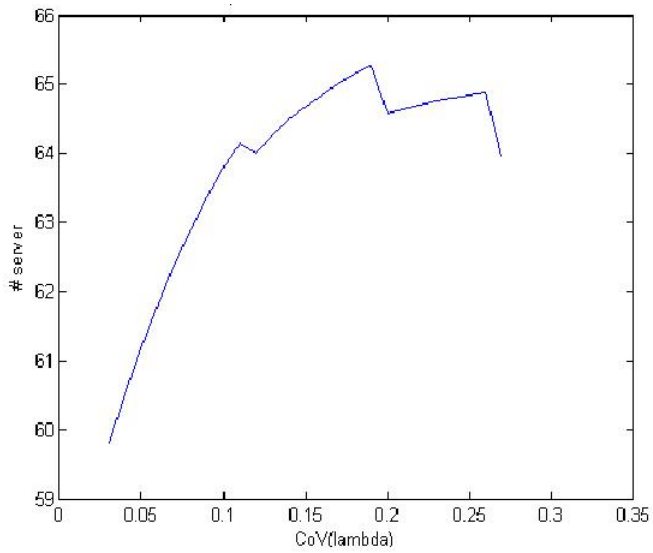

Figure 5.8: $x_1^*(n)$ vs $COV(\Lambda_1)$ for Probability of Waiting Model at $\frac{c^+-c}{c^+-c^-} = \frac{2}{5}$

5.2.4 Two-stage Model with Constraint on Probability of Abandonment

If we consider a single queueing system with abandonment, then a commonly used QoS constraint uses the probability of abandonment. We apply model (5.4) to a $M/M/n+M$ queueing system with a QoS constraint on the probability of waiting. In particular, we have

$$\min_{x_1 \geq 0} \quad cx_1 + \mathbb{E}_N h(x_1, N), \tag{5.10a}$$

$$\text{where} \quad h(x_1, N|_{N=n}) = \min_{x_2 \geq 0} \quad c^+(x_2 - x_1)^+ - c^-(x_1 - x_2)^+ \tag{5.10b}$$

$$\text{s.t.} \quad \mathbb{P}_{\Lambda_2|_{N=n}} \left(\mathbb{P}(\textit{abandonment} \mid x_2, \Lambda_2) < \delta \right) \geq 1 - \epsilon. \tag{5.10c}$$

In (5.10), $\mathbb{P}(\textit{abandonment})$ is calculated by the continuous Erlang-A formula which is given in (2.15) in Chapter 2.

As before due to the complexity of the formula for $\mathbb{P}(\textit{abandonment} \mid x_2, \Lambda_2)$, once we obtain n^* from step 1, it is not as easy as it is for the utilization constraint model to evaluate $x_2^*(n^*)$. Again, a line search is required.

In the next set of experiments, we solved (5.10) with the same set of values on α and β as in the experiments for solving (5.5). The ϵ and δ in (5.10) are both set to be 0.05. The abandonment rate θ is set to be 5. The lower and upper bounds of the line search are set to be 1 and 120, and the tolerance level of the line search is set to be 0.01. As in the experiment for the model with waiting probability constraint, in this experiment line searches on x_2 are needed in step 2 of the algorithm. We conducted the experiments again in MATLAB 7.11 (64 bit), and it took 946.41 seconds for MATLAB to

finish the experiments. These experiments demonstrate that our algorithm can efficiently solve the two-stage problem for more complex QoS measures using line searches. Although the solution times are obviously much greater than the times needed with the simple utilization metric, they are still quite reasonable.

Chapter 6

Conclusions and Future Directions

6.1 Summary

The service center industry is expanding rapidly, both in terms of its workforce and its economic scope. Today, most organizations have service centers as their major customer-facing channel, either internally managed or outsourced. Although the service center business has become more technology-intensive as new technologies emerge, most of its operating costs are still devoted to human resources. The statistics show that 60-80% (Aksin et al. [1]) of the over $300 billion (Gilson and Khandelwal [19]) in annual service center expenditures involve staffing costs. This gives rise to an intense research interest in service center staffing problems. Queueing networks are useful tools for analyzing complex stochastic service center systems. Furthermore, in practice, key model parameters, such as the arrival rates, are uncertain. In our work, we consider a doubly stochastic service center, meaning that in addition to inter-arrival times and service times being random, we model arrival rates as also being random. Our goal is to analyze the structural properties of the service center staffing problem in the presence of arrival-rate uncertainty. We further develop tools for optimizing staffing levels, using stochastic programming techniques.

In Chapter 2, we introduce the mathematical background for our work. We review two widely applied queueing models for service center problems, the Erlang-C model and the Erlang-A model, which are the basic models we use in our work. We also introduce approximations to the Erlang-C formula, which we apply in modeling the staffing problems. The mathematical proofs for the properties, such as uniform convergence of the approximations to the limiting formula, are given in this chapter. The properties of the approximations play a key role in proving the asymptotic optimality of our staffing policies in the later chapters.

In Chapter 3, we use an approximation of the service quality measurement in formulating a model that deals with the trade-off between staffing cost and service quality for the known arrival rate situation. In this chapter, we start by building a model for a single-class single-station service system, and then we extend the model to handle a multi-class multi-station system. We prove asymptotic optimality of solutions for our approximate model for both the single-class single-station system and the multi-class multi-station system.

We also focus on doubly stochastic service center systems; that is, we focus on solving large-scale service center staffing problems when the arrival rates are uncertain in addition to the inherent randomness of the system's inter-arrival times and service times. This brings the modeling closer to reality. In Chapter 4, we provide a solution procedure for solving a staffing problem for a doubly stochastic service center system. We consider a decision making scheme in which we must select staffing levels before observing the arrival rates.

We assume that the decision maker has distributional information about the arrival rates at the time of decision making. In the presence of arrival rate uncertainty, the decision maker's goal is to minimize the staffing cost, while ensuring the QoS achieves a given level. We show that as the system scales large in size, there exists at most one *key* scenario under which the probability of waiting converges to a non-trivial value, i.e., a value strictly between 0 and 1. In any other scenario, the staffing level is either over- or under-loaded in any other scenario as the size of the system grows to infinity.

Exploiting the notion of the key scenario, we propose a two-step solution procedure for the staffing problem with random arrival rates. In the first step, we use the desired QoS level to identify the key scenario corresponding to the optimal staffing level. After finding the key scenario, the random arrival-rate model reduces to a deterministic arrival-rate model. In the second step, we solve the resulting model, with deterministic arrival rate, by using our approximation model proposed in Chapter 3. The approximate optimal staffing level obtained in this procedure asymptotically converges to the true optimal staffing level for the random arrival-rate problem.

In Chapter 4, we focus on the staffing policy over a single decision time period in the presence of random arrival rates. In contrast, in Chapter 5, we build a two-stage stochastic program with recourse to analyze the relationship between the staffing decisions over two adjacent time periods. The problem considered in this chapter is a single-class single-station service center with random arrival rate. A Bayesian update is applied to the arrival rate in the

second time period once the new observations arrive during the first time period. The model integrates arrival-rate updates and dependence in staffing decisions over two contiguous time periods. The model minimizes the first stage staffing cost and the expected second stage staffing cost while satisfying a service quality constraint on the second stage operation. The Bayesian update yields the second-stage arrival-rate distribution based on the first-stage arrival-rate distribution and the observations in the first stage. The second-stage distribution is used in the constraint on the second stage service quality. After reformulation, we show that we can rewrite our two-stage model as a newsvendor model. We provide an algorithm that solves the two-stage staffing problem under some commonly used QoS constraints.

6.2 Future Work

In Chapter 3 and Chapter 4, when we consider multi-class multi-station systems, we assume there are only dedicated service agents. Our systems have no cross-trained agents. Also in the current scope of our research, we limit attention to the staffing decision at a service center. Two natural ways to extend our approach is to model heterogeneous service agents and to address the scheduling aspect of the problem. In practice, it is very likely that some of the service agents are cross-trained. Cross-trained agents can handle more than one kind of service request. Facing such a situation, after the staffing decision has been made, service center managers must then decide how to allocate the agents to service different types of requests. In practice, the

staffing decision on how many service agents are needed is made first and then scheduling decisions are made, in a hierarchical manner. However, there is potential benefit from integrating the staffing and scheduling decisions. We see our staffing models as ideal candidates to serve as submodels in such an integrated model.

Another way to extend our work in Chapter 3 and Chapter 4 is that, in the current setting, the QoS constraint we use in the model is a constraint on the probability a customer must wait. It would be interesting to see whether similar approximation models with other types of service quality measurement metrics, such as the probability of abandonment in an Erlang-A system, still lead to asymptotically optimal staffing policies.

In Chapter 5, we show that our model and algorithm applies to models with almost all major quality measurement metrics, such as utilization, the probability that an arriving customer waits in an Erlang-C system, and the probability a customer abandons in an Erlang-A system. However, in Chapter 5, we only consider problems with a single class and single station. Extending the model and algorithm to multi-class multi-station systems will be a challenging and interesting research direction for the future.

Bibliography

[1] Z. Aksin, M. Armony, and V. Mehrotra. The modern call-center: A multi-disciplinary perspective on operations management research. *Production and Operations Management*, 16(6):665–688, 2007.

[2] B. Andrews and S. Cunningham. LL Bean improves call-center forecasting. *Interfaces*, 25(6):1–13, 1995.

[3] M. Armony. Dynamic routing in large-scale service systems with heterogeneous servers. *Queueing Systems*, 51(3):287–329, 2005.

[4] R. Atar. Scheduling control for queueing systems with many servers: Asymptotic optimality in heavy traffic. *The Annals of Applied Probability*, 15(4):2606–2650, 2005.

[5] A. N. Avramidis, A. Deslauriers, and P. L'Ecuyer. Modeling daily arrivals to a telephone call center. *Management Science*, 50(7):896–908, 2004.

[6] A. Bassamboo, J. M. Harrison, and A. Zeevi. Dynamic routing and admission control in high-volume service systems: Asymptotic analysis via multi-scale fluid limits. *Queueing Systems*, 51(3-4):249–285, 2005.

[7] A. Bassamboo and A. Zeevi. On a data-driven method for staffing large call centers. *Operations Research*, 57(3):714–726, 2009.

[8] L. Bianchi, J. Jarrett, and R. C. Hanumara. Improving forecasting for telemarketing centers by ARIMA modeling with intervention. *International Journal of Forecasting*, 14(4):497 – 504, 1998.

[9] S. Borst, A. Mandelbaum, and M. I. Reiman. Dimensioning large call centers. *Operations Research*, 52(1):17–34, 2004.

[10] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association*, 100(469):36–50, 2005.

[11] L. Brown and H. Shen. Analysis of service times for a bank call center data. Technical report, Wharton School Center for Financial Institutions, University of Pennsylvania, 2002.

[12] H. E. Buchanan and T. H. Hildebrandt. Note on the convergence of a sequence of functions of a certain type. *Annals of Mathematics*, 9(2):123–126, 1908.

[13] B. Büke, J. J. Hasenbein, and D. P. Morton. Minimizing makespan in a multiclass fluid network with parameter uncertainty. *Probability in the Engineering and Informational Sciences*, 23(3):457–480, 2009.

[14] Datamonitor. http://www.datamonitor.com.

[15] A. Erlang. The theory of probability and telephone conversations. *Nyt Tidsskrift for Matematik*, B(20):33–39, 1909.

[16] A. Erlang. Solutions of some problems in the theory of probabilities of significance in automatic telephone exchanges. *Electroteknikeren*, 13:5–13, 1917.

[17] N. Gans, G. Koole, and A. Mandelbaum. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing and Service Operations Management*, 5(2):79–141, 2003.

[18] N. Gans, H. Shen, Y. P. Zhou, N. Korolev, A. McCord, and H. Ristock. Parametric stochastic programming models for call-center workforce scheduling. 2009. Working paper.

[19] K. A. Gilson and D. K. Khandelwal. Getting more from call centers. *The McKinsey Quarterly, http://www.mckinseyquarterly.com*, 2005.

[20] I. Gurvich, M. Armony, and A. Mandelbaum. Service level differentiation in call centers with fully flexible servers. *Management Science*, 54(2):279–294, 2008.

[21] I. Gurvich, J. Luedtke, and T. Tezcan. Staffing call centers with uncertain demand forecasts: A chance-constrained optimization approach. *Management Science*, 56(7):1093–1115, 2010.

[22] S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29(3):567–588, 1981.

[23] J. M. Harrison and A. Zeevi. A method for staffing large call centers based on stochastic fluid models. *Manufacturing and Service Operations Management*, 7(1):20–36, 2005.

[24] A. A. Jagers and E. A. Van Doorn. On the continued Erlang loss function. *Operations Research Letters*, 5(1):43–46, 1986.

[25] A. J. E. M. Janssen, J. S. H. Van Leeuwaarden, and B. Zwart. Refining square root safety staffing by expanding Erlang C. *Operations Research*, 59(6):1512–1522, 2011.

[26] G. Jongbloed and G. Koole. Managing uncertainty in call centers using poisson mixtures. *Applied Stochastic Models in Business and Industry*, 17(4):307–318, 2001.

[27] S. Maman. Uncertainty in the demand for service: The case of call and emergency departments. Master's thesis, Technion-Israeli Institute of Technology, 2009.

[28] A. Mandelbaum and S. Zeltyn. Service engineering in action: The Palm/Erlang-A queue, with applications to call centers. In *Advances in Services Innovations*, pages 17–45. Springer-Verlag, Berlin, 2007.

[29] C. Palm. Intensitätsschwankungen im fernsprechverkehr. *Ericsson Technics*, 44:189 pp, 1943.

[30] C. Palm. Research on telephone traffic carried by full availability groups. *Tele*, 1:107, 1957.

[31] T. K. Ralphs, M. J. Saltzman, and M. M. Wiecek. An improved algorithm for solving biobjective integer programs. *Annals of Operations Research*, 147:43–70, 2006.

[32] J. Riordan. *Stochastic Service Systems*. Wiley, New York, 1962.

[33] T. R. Robbins and T. P. Harrison. A stochastic programming model for scheduling call centers with global service level agreements. *European Journal of Operational Research*, 207(3):1608–1619, 2010.

[34] H. Shen and J. Z. Huang. Analysis of call centre arrival data using singular value decomposition. *Applied Stochastic Models in Business and Industry*, 21(3):251–263, 2005.

[35] H. Shen and J. Z. Huang. Forecasting time series of inhomogeneous Poisson processes with application to call center workforce mangement. *Annals of Applied Statistics*, 2(2):601–623, 2008.

[36] H. Shen and J. Z. Huang. Interday forecasting and intraday updating of call center arrivals. *Manufacturing and Service Operations Management*, 10(3):391–410, 2008.

[37] H. Shen, J. Z. Huang, and C. Lee. Forecasting and dynamic updating of uncertain arrival rates to a call center. *Service Operations and Logistics, and Informatics*, pages 1–6, 2007.

[38] R. Soyer and M. M. Tarimcilar. Modeling and analysis of call center arrival data: A Bayesian approach. *Management Science*, 54(2):266–278, 2008.

[39] J. W. Taylor. Density forecasting of intraday call center arrivals using models based on exponential smoothing. *Management Science*. To appear.

[40] J. W. Taylor. Short-term electricity demand forecasting using double seasonal exponential smoothing. *Journal of the Operational Research Society*, 54(8):799–805, 2003.

[41] W. Tych, D. J. Pedregal, P. C. Young, and J. Davies. An unobserved component model for multi-rate forecasting of telephone call demand: The design of a forecasting support system. *International Journal of Forecasting*, 18(4):673–695, 2002.

[42] J. Weinberg, L. D. Brown, and J. R. Stroud. Bayesian forecasting of an inhomogeneous Poisson process with applications to call center data. *Journal of the American Statistical Association*, 102(480):1185–1198, 2007.

Vita

Jing Zan was born in Beijing, China, in 1983. She received a B.S. in Mathematics in 2005 from Shandong Univeresity, Shandong, China, and a M.S. in Operations Research and Industrial Engineering in 2008 from the University of Texas at Austin, Texas, USA. She has been working on her doctorate at the University of Texas at Austin since 2008.

Permanent address: 1618 West Sixth Street
Apt A
Austin, TX 78703

This dissertation was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.